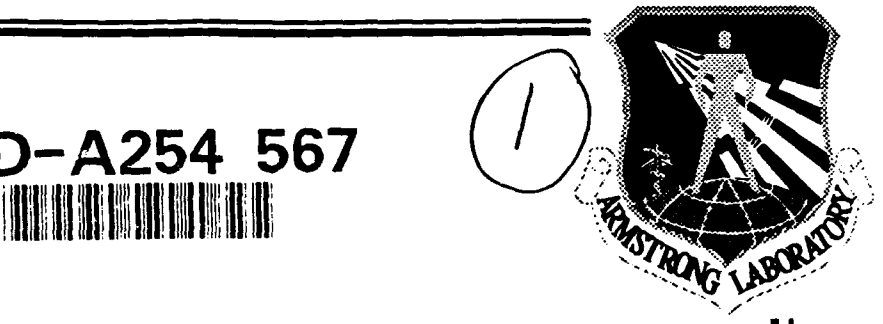
AL-CR-1992-0003



ARMSTRONG LABORATORY

EXAMINING OPERATIONAL MEASURES OF PERFORMANCE AND DEVELOPING METHODS FOR DETERMINING COMPETENCY LEVELS FOR THE AIR FORCE JOB PERFORMANCE MEASUREMENT SYSTEM: LITERATURE REVIEW AND METHODS

Carolyn Hill Fotouhi
Gregory P. Mosher
Rodney A. McCloy

Human Resources Research Organization
1100 South Washington Street
Alexandria, VA 22314

HUMAN RESOURCES DIRECTORATE
TECHNICAL TRAINING RESEARCH DIVISION
Brooks Air Force Base, TX 78235-5000

July 1992

Final Contractor Report for Period January 1990 – July 1990

Approved for public release; distribution is unlimited.

AIR FORCE MATERIEL COMMAND
BROOKS AIR FORCE BASE, TEXAS 78235-5000

NOTICES

This contractor report is published as received and has not been edited by the technical editing staff of the Armstrong Laboratory.

Publication of this report does not constitute approval or disapproval of the ideas or findings. It is published in the interest of scientific and technical information (STINFO) exchange.

When Government drawings, specifications, or other data are used for any purpose other than in connection with a definitely Government-related procurement, the United States Government incurs no responsibility or any obligation whatsoever. The fact that the Government may have formulated or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication, or otherwise in any manner construed, as licensing the holder, or any other person or corporation; or as conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

The Office of Public Affairs has reviewed this report, and it is releasable to the National Technical Information Service, where it will be available to the general public, including foreign nationals.

This report has been reviewed and is approved for publication.

Donald L. Harville

DONALD L. HARVILLE
Project Scientist

HENDRICK W. RUCK, Technical Director
Technical Training Research Division

RODGER D. BALLENTINE, Colonel, USAF
Chief, Technical Training Research Division

July 1992 Final – January 1990 – July 1990

Examining Operational Measures of Performance and Development Methods for Determining Competency Levels for the Air Force Job Performance Measurement System: Literature Review and Methods

C - M67004-89-D-0010
PE - 63227F
PR - 2922
TA - 01
WU - 05

Carolyn H. Fotouhi
Rodney A. McCloy
Gregory P. Mosher

Human Resources Research Organization
1100 South Washington Street
Alexandria, VA 22314

Armstrong Laboratory
Human Resources Directorate
Technical Training Research Division
Brooks Air Force Base, TX 78235-5000

AL-CR-1992-0003

Armstrong Laboratory Technical Monitor: Donald L. Harville, (512) 536-2932.

The Job Performance Measurement (JPM) data needs to be interpreted in terms of job competency levels. The literature on determining standards for job competency levels was reviewed. Various item-based methods (i.e., Nedelsky, Angoff, Ebel, and Jaeger methods) and examinee-based methods (i.e., contrasting groups and borderline groups methods) were reviewed and compared. An examinee-based method using archival JPM proficiency ratings, by work sample test administrators, was proposed. An item-based method using subject matter experts and normative data was proposed. Models of utility, used to evaluate the consequences of standards, were discussed. An annotated bibliography of the standard setting literature was included as an appendix.

Job competence
Performance appraisal
Performance measures
Standard setting
Work samples
Work standards

114

Unclassified Unclassified Unclassified UL

TABLE OF CONTENTS

List of Tables

List of Figures

Project Overview

The Job Performance Measurement/Enlistment Standards (JPM) Project, initiated in 1980, is a joint-Service research program exploring methods for assessing the job performance capability of enlisted personnel. As part of this project, the Air Force Human Resources Laboratory (AFHRL) has developed and administered different types of performance measures to first term airmen in various occupations. These efforts have led to the conclusion that hands-on testing is the most valid form of performance measurement. However, hands-on testing is not always feasible or practical because of cost, safety, and other factors. Due to its impracticality, hands-on tests can not be operationalized in most situations, therefore, they will serve as benchmarks by which surrogate measures can be assessed. Four types of job performance measures (i.e., hands-on tests, interview tests, rating forms, and knowledge tests) were developed and administered to selected enlisted specialties.

With these tasks accomplished, the question remains of how these newly-developed measures will be used operationally. The most appropriate and effective uses of this technology must still be investigated. Our investigation will result in recommendations for the applications of these performance measures.

Project Objective

The fate of the Job Performance Measurement System (JPMS), developed for the JPM Project has not yet been determined. Work needs to be completed to ascertain the most cost-effective and germane uses of the available measurement methods for airmen. The objective of this study is to refine the JPM project in light of the measures currently used and to investigate the possible use of levels of competence for attaching meaning to job performance scores. This effort is twofold: (a) Investigate the current methods by which data is gathered for job performance purposes, evaluate the usefulness of these methods, and make recommendations on the extent to which the JPM measures can be used, either as supplements or as a means of validating existing methods; and (b) Explore competency levels for different criterion measures and quantify the advantages and disadvantages of those competency levels.

Plan Summary

This project is divided into two phases. The first phase involves analyzing competency levels, determining cutoff scores, quantifying their advantages and disadvantages, and evaluating and quantifying the resources needed (e.g., Subject Matter Experts [SMEs], job analysis information, etc.). The second phase consists of the evaluation of the existing criteria, performance measures, and other indicators of performance.

Proficiency Measurement

As stated in the project overview, the JPM Project of the Armed Services was established to examine the feasibility of measuring job performance and to link enlistment standards to job performance (Green & Wigdor, 1988). Competency and performance are terms which are very much intertwined within this informal technical report. As stated in the evaluation of the JPM Project (Green & Wigdor, 1986), "to be really useful in the central matter of setting standards and allocating recruits among job specialties, the project's primary measurement goal should be to supply performance scores with some absolute meaning, i.e., to measure individuals' proficiency with reference to the whole job. This we have designated as a competence approach" (p. 55). By ascribing meaning to the performance measures (e.g., hands-on performance tests [HOPT]) and the resulting performance scores, and by referencing them to an external scale of job requirements, job competency can be more easily determined. Three different methods of assessing competency will be compared and contrasted. First, however, we must define what we mean by competence to provide a frame of reference for the resulting comparisons.

Competence Defined

In today's literature there are as many definitions of competence as there are ways of measuring it. Webster defines competence as "the quality of being functionally adequate or of having sufficient knowledge, judgement, skill, or strength." Determining what is required for a given job, while necessary, is only the beginning of the process of measuring competence. When it comes to measuring levels of competence or proficiency and assessing the value of individuals to a total system output, a more precise definition is needed.

Competence has been referred to in many contexts, and in almost every case, competence is viewed as a dichotomous variable (i.e., competent vs. incompetent). In the legal profession, competence refers to whether a person is able to stand trial. In the educational literature, the term competence has been used to assess whether persons have attained a certain level of proficiency by comparing performance against some criterion or cutoff. Recently competence has been viewed in a more cognitive sense, "as an examinee's actual level of cognitive functioning, if performance impediments were removed or eliminated" (Dillon & Stevenson-Hicks, 1983). Glaser, Lesgold, and Gott (1986) also suggest a more cognitive approach to competence; the determinants of competence may not always be revealed by the surface characteristics of either the worker's performance or the environment in which that performance takes place. Glaser et al. analyzed what skilled workers do to perform at a higher level of proficiency. They point out that detailed technical data are not committed to memory, but "streamlined mental representations" of the workings of a system replace the cut-and-dry "how to" of a system, therefore simplifying the job.

Wood and Power (1987) draw an interesting distinction between competence and performance: "Competence refers to what a person knows and can do under ideal circumstances, whereas performance refers to what is actually done under existing circumstances. . . . Developed competence is to be conceived of and

assessed as a continuous variable reflecting various degrees of integration of knowledge and skill, of understanding and proficiency" (p. 415).

In line with the stated goals of the JPM Project is the view that to derive the most useful information from the performance measures, an absolute meaning must be attached to a performance score. Such an absolute meaning can be established with the measurement of competence or job proficiency. As Glaser (1963) wrote,

> Underlying the concept of achievement measurement is the notion of a continuum of knowledge acquisition ranging from no proficiency at all to perfect performance. An individual's achievement level falls at some point on the continuum as indicated by the behaviors he displays during testing. The degree to which his achievement resembles desired performance at any specified level is assessed by criterion-referenced measures of achievement or proficiency. The standard against which a student's performance is compared when measured in this manner is the behavior which defines each point along the achievement continuum. The term "criterion," when used in this way, does not necessarily refer to final end-of-course behavior. Criterion levels can be established at any point in instruction as to the adequacy of an individual's performance. The point is that the specific behaviors implied at each level of proficiency can be identified and used to describe the specific tasks a student must be capable of performing before he achieves one of these knowledge levels. It is in this sense that measures of proficiency can be criterion-referenced.
>
> Along such a continuum of attainment, a student's score on a criterion-referenced measure provides explicit information as to what the individual can or cannot do. Criterion-referenced measures indicate the content of the behavioral repertory, and the correspondence between what an individual does and the underlying continuum of achievement. Measures which assess student achievement in terms of a criterion standard thus provide information as to the degree of competence attained by a particular student which is independent of reference to the performance of others (p. 519).

Glaser (1963), and Popham and Husek (1969) were among the first to popularize the field of criterion-referenced testing (Hambleton & Swaminathan, 1978). The differences between criterion- and norm-referenced measurement deserve attention.

Norm-referenced vs. Criterion-referenced Measurement

In both norm-referenced (NR) and criterion-referenced (CR) measurement, samples of test items are drawn from a population of items representing the domain of task performance. Yet there are some basic distinctions to be made between the two processes. NR measures are employed to ascertain an individual's proficiency in relation to other individuals, thus providing a ranking. CR measures ascertain an individual's status with respect to a criterion (i.e., performance standard) and a well defined behavior domain. It is because the individual is compared with some established criterion, rather than other individuals, that these measures are described as criterion-referenced (Popham & Husek, 1969). NR measures yield information regarding individuals whereas CR

measures yield information on individuals and treatments (e.g., instructional programs). NR measures are particularly useful when a quota is involved, or selection of top candidates is an issue. Although information as to the relative standing of individuals as compared to their system counterparts reveals useful information, our approach to competency measurement warrants a system which yields information as to the absolute standing in terms of proficiency of job content. Therefore a criterion-referenced approach will be taken.

A Review of Issues and Procedures for Setting Performance Standards and Determining Job Proficiency

Obtaining individual proficiency information from work sample tests and referencing that information to a continuum of job proficiency is done with what Guion (1978) calls "content-referenced" testing. Guion (1978) states,

> Some form of content-referenced scale is usually necessary to provide adequate meaning for a work sample test. At least three kinds of content-referenced scales can be devised.
>
> 1. Occasionally, a group of expert judges, considering the examination in detail, will arrive at a system for establishing an arbitrary cutting point or standard above which mastery may be claimed. Where such a standard is established, scores can be interpreted in terms of linear distances from that standard point. This can be a useful scale, but its value depends on how widely the standard is accepted.
>
> 2. A priori scaling can provide a basic reference scale. If a subset of test components or items form a scale, then total scores can be interpreted with reference to that scale of selected items.
>
> 3. If latent trait analysis is used, the test can be scored on the basis of maximum likelihood estimates or other estimates along a "sample-free" scale of underlying latent ability (p. 8).

The three methods proposed by Guion will be the basis for our discussion of standard setting, competency and related research. For ease of reference our discussion will follow the order of presentation in Guion (1978).

Standard Setting Methods

Standard setting procedures can be divided into two categories: (a) item-based and (b) examinee-based. Item-based methods require that raters make judgments regarding the proportion of minimally competent individuals who would correctly answer each test item. Proportions are aggregated across items and judges to form a "percent correct" standard. Examinee-based methods require that raters identify competent and noncompetent or borderline competent individuals who are then administered the test. Standards are then set based on the data obtained from the test administration. As can be seen from these general descriptions, all standard setting procedures require some subjective judgments. With item-based approaches, judges are required to make decisions about the test

items. With examinee-based approaches, decisions are required concerning individuals who will be administered the test.

There has been much discussion in the literature regarding the subjectivity inherent in all standard setting procedures. Glass (1978a, 1978b) traces the evolution of the notion of performance standards and concludes that all standard setting procedures are so arbitrary that they are worthless. He argues that standards cannot be set without consideration of their consequences, yet he contends that procedures designed to evaluate the consequences of various cutoffs are too arbitrary to be useful. Block (1978) and Popham (1978), on the other hand, argue that although standard setting requires some subjective decisions these decisions are not made in a vacuum. Block notes that early standard setting procedures possessed some logic and that more recent procedures, while imperfect, have attempted to improve on that logic. In addition to improving current standard setting methodology, Popham asserts that the task at hand for both applied and theoretical researchers is to determine the most appropriate procedure.

Most literature on standard setting comes from education, where the primary concern is setting minimum competency standards on written, multiple choice tests. Although Jaeger and Keller-McNulty (1986) suggest that the methods used to establish standards on written tests can be modified for use with performance tests, essentially no research has attempted to apply these techniques to performance tests. The following is a brief review of traditional standard setting procedures. More comprehensive reviews can be found by Pulakos, Wise, Arabian, Heon, and Delaplane (1989) and Jaeger and Keller-McNulty (1986).

Item-Based Methods

In setting performance standards, item-based methods are more widely used than examinee-based methods. Perhaps because they are so widely used, the item-based methods have become synonymous with the names of their respective developers. The four item-based standard setting methods discussed in the following section are (a) Nedelsky, (b) Angoff, (c) Ebel, and (d) Jaeger.

Nedelsky method. The Nedelsky method requires a multiple choice test format. For each item, judges identify the distractors that a minimally competent individual would readily eliminate as incorrect. A minimum passing level (MPL) is then calculated for each item. The MPL is equal to the reciprocal of the remaining response options, after eliminating easily identifiable incorrect options. For example, if a judge identifies two of five distractors as easily eliminated by a minimally competent individual, the MPL is 1 divided by 3 or .33. Thus, the MPL is calculated for each item for each judge. A cutoff score is obtained for each judge by summing the MPLs across items for that judge. A standard for the test is obtained by averaging MPLs across judges.

To avoid classifying as incompetent, an examinee whose true performance is just equal to the test standard, solely as a result of measurement error, Nedelsky (as cited in Jaeger & Keller-McNulty, 1986) recommends a downward adjustment of the initial standard. Working from several assumptions, the adjustment requires reducing the initial standard by one or more standard deviations of the distribution of MPLs obtained from the sample of judges.

The Nedelsky method is the most widely used method for setting standards for professional certification and licensure exams (Livingston & Kastrinos, 1982). However, there are several disadvantages to the method. The first disadvantage lies in the assumptions regarding examinee decision making processes. Once obviously incorrect options are identified, examinees are assumed to have no information, however partial it may be, on which to select from among the remaining response options. Therefore, it is assumed that examinees randomly choose among response options they cannot clearly identify as incorrect. In reality, single test items are not presented in a vacuum as these assumptions lead one to believe. Information from one item may be, and often is, used to help answer another item. The Nedelsky method does not account for this. Thus, resulting standards may be more lenient than intended.

Poggio (1984) summarizes several shortfalls with the method based on successive implementations with the Kansas minimum competency testing program. He found that raters were often confused by the method, and as a result, they reported not being confident in their judgments. Raters also tended to be careless in studying items and often designated the correct response as a viable distractor. Because the method is confusing, highly trained raters are required. While it is imperative that raters be experts in the area in which standards are to be set, it seems wasteful to spend extra time and resources training them on how to use a particular standard setting method when another method will work effectively without such extra training.

A primary disadvantage of the Nedelsky method given the goals of the JPMS is that it can be used to set standards only on multiple choice tests. Therefore, it could be used to establish standards on the written job knowledge tests but not for any other job performance measures (i.e., hands-on tests, interview tests, and rating scales).

Angoff method. The Angoff method asks raters to think of a group of minimally competent individuals rather than only one person. Raters estimate the percentage of minimally competent individuals who would be able to answer each item correctly. The cutoff score for a particular rater is the sum of his or her percentages across items. The test standard is the average of cutoff scores across raters. Thus, the percentage of minimally competent individuals passing an item is converted to the percentage of items that should be passed by minimally competent individuals.

Compared to other standard setting procedures, the Angoff method is the most straightforward and the easiest to implement. Raters have essentially no problem understanding the task they are to perform.

One disadvantage is the amount of variability in the standards provided by the Angoff method (Poggio, Glassnap, & Eros, 1981). Variability is particularly a problem when only a few raters are used as is the case in most workshop settings. Jaeger and Busch (1984) used an iterative approach and normative data in an attempt to reduce the variability in ratings obtained via the Angoff method. Raters first provided independent ratings. After the presentation of normative data and a discussion period, raters were allowed to independently reconsider their original ratings. While the mean standard did not change significantly, variability in the ratings was reduced. Using only

an iterative procedure and no normative data, Norcini, Lipner, Langdon, and Strecker (1987) also found a reduction in standard variability. No research could be found that examined the advantages and disadvantages of an iterative approach only, normative data only, or a combination.

The method is appropriate only for dichotomously scored items (Pulakos et al., 1989). However, the method could be modified for tests composed of continuous scale measures (e.g., assessments based on rating scales) by asking SMEs to estimate the most likely, or average, rating for minimally competent persons. Averaging these ratings across the performance measures provides a cutoff recommendation for each SME. SME recommendations can then be averaged to give an overall test standard.

Ebel method. The Ebel method requires subject matter experts (SMEs) to classify test items on two dimensions: (a) difficulty and (b) relevance. Ebel suggested three levels of difficulty (easy, medium, and hard) and four levels of relevance (essential, important, acceptable, and questionable). However, the dimensions and number of levels can be changed without altering the basic method. After considering each item on the two dimensions, SMEs working independently allocate each item to 1 of the 12 cells formed by the 3 (difficulty) x 4 (relevance) matrix. For example, item 1 might be judged to be "easy" and of "questionable relevance"; item 2 might be judged to be "hard" and "essential"; etc. Working as a group, SMEs then decide the percentage of minimally competent examinees who would be able to correctly answer items in each of the 12 cells. Percentages are assigned to cells without regard to the particular items in each cell. For example, 90% of minimally competent examinees might be expected to correctly answer "easy and essential" items; 20% might be expected to correctly answer "hard and questionable" items; etc.

For each SME, the number of items in a particular cell is multiplied by the percentage assigned to that cell. These products are summed across cells to yield a cutoff score for each SME. The average cutoff score across SMEs becomes the test standard.

SMEs find the traditional Ebel method easy to understand and implement. However, it is time-consuming. Boredom and fatigue may become a problem, especially if the test contains many items, and setting multiple cutoff scores exacerbates the problem. Other disadvantages are associated with the method. Poggio (1984) found that many SMEs were troubled by the "questionable" label on the relevance dimension. Because the dimensions and number of levels within a dimension are irrelevant to the basic method, this problem can easily be eliminated. Another disadvantage is that the Ebel method consistently results in stricter standards than other standard setting methods (Andrew & Hecht, 1976; Poggio, 1984; Skakun & Kling, 1980).

Unmodified, the Ebel method is restricted to use with dichotomously scored items (Pulakos et al., 1989). Similar to the Angoff method, the Ebel method could be modified for tests composed of items measured on a continuous scale. The original Ebel questions essentially ask for the average score (i.e., percent passed) of minimally competent persons. A modified version would be to ask SMEs to estimate the most likely rating, or average rating, for the measures within each of the matrix cells. Averaging these ratings across the cells, weighted

for the number of measures in each cell, provides a cutoff recommendation for each SME. SME recommendations can then be averaged to give an overall test standard.

In the suggested modifications, the Ebel method differs from the Angoff method only in that SMEs rate categories of items instead of individual items, thus, the Ebel method requires that items be categorized. The modifications suggest the same question for both methods: What is the average score for minimally competent persons?

Jaeger method. Poggio (1984) points out that many raters have difficulty determining the percentage of examinees who should correctly answer each item. The Jaeger method circumvents that problem by having raters answer a yes/no question. Instead of trying to estimate the performance of minimally competent individuals, judges are asked to consider the following question: "Should every examinee in the population of those who receive favorable action on the decision that underlies use of the test be able to answer the test item correctly?" (Jaeger & Keller-McNulty, 1986, p. 14). In other words, should every person who is at least a minimally competent examinee be able to answer this item correctly? A "yes" response is scored as 1, and a "no" response is scored as 0.

In the first phase, judges independently answer the above question for each test item. An initial cutoff score for each judge is calculated by summing his or her "yes" responses across items. An initial test standard is determined by computing the median cutoff score across judges. While the Nedelsky and Angoff methods have been modified to include the use of normative data and an iterative approach (Cross, Impara, Frary, & Jaeger, 1984; Koffler, 1980; Norcini et al., 1987), the Jaeger method prescribes these conditions at a minimum.

For the iterative approach, the percentage of examinees who actually answered each item correctly on a recent administration of the test is presented after SMEs make their initial judgments. Upon reviewing the data, judges are asked to reconsider their recommendations and again independently answer the same question for each item. A second cutoff score is computed for each judge, and a second test standard is computed for the entire group.

In preparation for the final rating phase, more normative data is provided. Specifically, given the group's second standard, judges are told the percentage of examinees who would have failed the test on a recent administration. The distribution of cutoff scores recommended by fellow judges during the second phase is also presented. Judges once again answer the same yes/no question. Using the same computational procedure, a final standard is calculated for each judge and for the group. The median standard for the group becomes the test standard.

The Jaeger method inherently requires that judgments be made in a workshop setting. The nature of the information presented and the ensuing discussion requires a skilled workshop leader. The advantages and disadvantages of a workshop setting depend upon the frequency with which standards are set and the standard setting experience of the raters. If standards are to be set frequently as with an ongoing minimum competency testing program, a workshop approach will quickly become expensive and time-consuming.

Perhaps the greatest disadvantage of the Jaeger method is that, like the traditional Ebel method, it is time-consuming. While fatigue and boredom may become a problem, it is not likely to be as pervasive as with the Ebel method. The Ebel method requires judges to consider items on two dimensions and little time is allotted to group discussion. Although raters answer the same question several times with the Jaeger method, only a simple yes or no answer is required, and more time is allotted to group discussion.

Finally, like the previously discussed methods, the Jaeger approach was also designed for use with dichotomously scored items. And like the other methods, the basic question put to the SMEs can be restructured to adapt the method to tests composed of continuously scored scales. In this case, the appropriate question could be stated as follows: What is the lowest score that should be observed among persons who receive favorable actions on the decision that underlies use of the test? Or simply, what is the lowest acceptable score?

The Jaeger method can be viewed as a combination of the item-based and examinee-based approaches to standard setting. Item-based approaches require decisions about test items, and examinee-based approaches require decisions about examinees. By using normative data, the Jaeger method requires decisions about test items in light of examinee performance on those items. Examinee-based methods are discussed next.

Examinee-Based Methods

A basic assumption underlying examinee-based approaches is that judges who are familiar with examinee performance in the knowledge, skill, and ability (KSA) being tested are capable of identifying individuals who are high in the KSA and those who are low. In other words, it is assumed that expert judges can conceptualize distinct levels of performance, and independent of data from the test in question, can identify individuals at each level. Not only are experts quite accurate in predicting the performance of individuals whom they know well, but also lay persons feel confident in those predictions. For example, in education where teachers serve as standard setting judges, parents readily accept the standards established via an examinee-based approach (Poggio, 1984). They often feel minimum competency testing is unnecessary because teachers can identify competent and non-competent students without using the test data.

Evaluations required of supervisors are similar to those required of teachers. In addition to formal evaluations, supervisors make informal assessments of subordinates who need remedial training, of those who are ready for additional responsibility, etc. In order to make these assessments, supervisors must be familiar with the KSAs required by the job as well as the performance of subordinates in regard to those KSAs. Furthermore, most of these assessments are made without the aid of test data.

A second assumption underlying examinee-based standard setting approaches is that most judges are more accustomed to making decisions about individuals than making decisions about test items. This is especially true of supervisors. As mentioned earlier, most supervisory decisions are made without relying on test data. In fact, supervisors rarely, if ever, administer formal tests to their

subordinates. Therefore, supervisors are even more likely than teachers to be more comfortable making decisions about individuals than about test items or tests. The two examinee-based approaches discussed in the following section are (a) Contrasting Groups method and (b) Borderline Group method.

Contrasting Groups method. According to the Contrasting Groups method, judges are asked to identify individuals who fall into one of two groups: competent vs. non-competent. Once the groups have been identified, the test is administered to them, and the distributions of scores are compared. The cutoff score is selected to maximally differentiate between the score distributions of the groups. The use of two groups results in a single test standard; however, two or more standards may be set by increasing the number of groups. For example if two standards are desired, individuals may be classified as competent, marginal, or non-competent.

One drawback of the Contrasting Groups method is the subjective process of identifying competent and non-competent individuals. To eliminate the subjective judgement, a modification of the Contrasting Groups method, administering the test to instructed and non-instructed individuals, is suggested. The assumption is that instructed individuals are competent and non-instructed individuals are non-competent. In this way, one omits the judgmental process of identifying competent and non-competent individuals.

Several methods for analyzing the distributions and selecting a standard have been proposed. The simplest method is to plot the distributions on a single graph. The score at which the distributions intersect is selected as the test standard. This method is applicable if the distributions are not coincident and if they overlap, especially if they overlap at a single, clear point. In reality, such a pattern rarely occurs. Fisher (as cited in Poggio et al., 1981) suggests several variations of statistical procedures for establishing standards which consider the shapes and relative variances of the distributions. If the groups have normal distributions and equal variances, the Linear Discriminant Function (LDF) is appropriate. If the distributions are normal and variances are unequal, the Quadratic Discriminant Function (QDF) is used. When the distributions are not normal, non-parametric analogs to the LDF and QDF (for equal and unequal variances, respectively) are appropriate.

Borderline Group method. In implementing the Borderline Group method, judges are asked to identify individuals who are "borderline" between competent and non-competent (i.e., they cannot be clearly identified as competent or non-competent). The test is administered to these individuals, and the resulting median test score for the group defines the test standard. Many of the advantages and disadvantages of the Contrasting Groups and Borderline Group methods are the same or very similar. Therefore, pros and cons of the two methods are discussed together.

To obtain accurate, unbiased standards with the Contrasting Groups and Borderline Group methods, it is imperative that individuals selected for testing be carefully identified and classified. Raters must consider only the KSAs covered by the test and classify individuals accordingly. For example, if the test covers reading comprehension, raters must classify individuals as competent, borderline, or non-competent on reading comprehension, not some other ability.

Thus, raters must be familiar with the performance of the individuals being classified. However, as the raters' familiarity with the individuals being classified increases, so does the probability of halo error. Fortunately, numerous studies (Borman, 1975; Latham, Wexley, & Pursell, 1975; Pulakos & Borman, 1985) have shown that rater training is effective in reducing halo error.

SMEs report few problems identifying competent and non-competent individuals, but many have difficulty identifying "borderline" individuals (Mills, 1983; Poggio, 1984). Mills concludes that examinees may be classified as "borderline" merely because SMEs lack sufficient information on which to base a decision.

One potential disadvantage of the examinee-based methods is that the cost of administering the test as a prerequisite for setting standards may not be feasible, especially for performance tests. While administering a written test may not be very expensive, performance tests often require more resources. It is possible to circumvent a separate, and potentially expensive, test administration by using data from the first test administration to establish performance standards. In this case, standards are not known prior to the first test administration. While this may be a less expensive solution, it is difficult to convince lay persons of the credibility of standards set in this fashion.

While examinee-based methods prescribe the classification of examinees prior to test administration, Cantor (1989) applied both the Contrasting Groups and Borderline Group procedures to archival data. Although the purpose of the study was to evaluate a previously established Ebel-derived standard, it is of interest because it is the only study to use examinee-based procedures to establish standards on archival data. Several criteria that were external to the test in question were identified and used to classify examinees as competent or non-competent. Although some classification errors resulted from partial information used to classify examinees, the methodology provides a less subjective means of classifying competent and non-competent examinees.

Aside from simplicity (Poggio, 1984), the primary advantage of the Contrasting Groups and Borderline Group methods is that they are more objective than the item-based methods. Once SMEs have identified competent, borderline, and non-competent individuals, the subjective phase is complete. For the Contrasting Groups method, decisions must be made regarding the proper use of statistics, but the characteristics of the score distributions will dictate the appropriate statistical analyses to be performed. While these methods are considered more objective by many researchers and practitioners, people who do not understand the statistical manipulations may be confused and doubt the validity of standards established through their use (Poggio, 1984).

Comparisons of Item-based and Examinee-based Methods

Studies to examine similarities and differences among standard setting methods of written tests have been conducted. However, any similarities and/or differences among standard setting procedures as applied to performance tests remain unknown. It is generally assumed that results obtained from comparisons of written tests are applicable to performance tests. Aside from some general

considerations, the consensus in the literature seems to be that the process itself is not as important as whether the standards are realistic (Buck, 1977) and whether the procedure is feasible given situational constraints such as financial and human resources, time available, appropriateness of the method for the type of test being studied, etc. (Hambleton, 1980).

In most research comparing standard setting methods, only item-based procedures are examined. Most such comparisons consider the Nedelsky method and one or more additional item-based procedures. Research results have consistently shown that the Nedelsky method produces the lowest and most unreliable standards (Brennan, & Lockwood, 1980; Cross et al., 1984; Halpin, & Halpin, 1987; Halpin, Sigmon, & Halpin, 1983). The Ebel method tends to produce the strictest standards (Poggio et al., 1981), and the standard produced may (Halpin, & Halpin; Poggio et al.) or may not (Andrew, & Hecht; Poggio, 1984) be highly reliable. The Angoff and Jaeger methods produce standards that typically fall somewhere between those produced by the Nedelsky and Ebel methods with a tendency for Jaeger standards to be stricter than Angoff standards (Cross et al.; Jaeger, & Keller-McNulty, 1986).

Few studies investigate examinee-based methods. In summarizing his findings across several years of standard setting for Kansas competency tests, Poggio (1984) found that standards produced by the Contrasting Groups and Borderline Group methods tend to be lower than those produced by the Angoff procedure. With one group of raters using different procedures, Mills (1983) found no differences in the standards set with the Angoff, Contrasting Groups, and Borderline Groups methods. Mills points out that although different methods may have produced different results, at least some of the discrepancies between methods probably have been due to differences between groups of judges.

In comparing the ease of implementation among methods, Poggio (1984) found the Angoff, Ebel, Contrasting Groups, and Borderline Group methods easily implementable and comprehensible. His research did not examine the Jaeger method, but he found that judges were confused by the Nedelsky method. In general, results taken across studies show that the Angoff method is the easiest to implement and that raters more readily comprehend the task they are to perform compared to other methods. Although previous research found that raters sometimes had difficulty identifying borderline individuals (Mills, 1983; Poggio et al., 1981), it is believed that an exhaustive definition including hypothetical examples can overcome this confusion.

When generalized across studies and standard setting procedures, the perception is that: (a) the Ebel method produces the highest standards, (b) the Nedelsky method produces the lowest, (c) the Angoff and Jaeger methods produce standards somewhere in the middle, and (d) most examinee-based methods are not feasible. Because of the disparity in standards established by the various procedures, many researchers recommend the use of several standard setting procedures to set performance standards (Halpin et al., 1983; Koffler, 1980).

Standard Setting Process

There are several issues concerning the standard setting process that are independent of the procedure used. Before deciding on the appropriate method, these issues warrant examination. They are discussed in the following sections.

Characteristics of judges. The identification and utilization of qualified experts is perhaps the most important consideration in any standard setting procedure. Research results in the field of education indicate that different groups of judges from a variety of backgrounds, if qualified, provide similar standards. In addition, standards are more readily accepted if they are set by qualified judges from a number of backgrounds (Andrew & Hecht, 1976; Jaeger, 1976).

Employing a variety of judgmental standard setting procedures, the U.S. Army's Synthetic Validity Project (Peterson, Owens-Kurtz, Hoffman, Arabian, & Whetzel, 1989) used NCOs and Officers from FORSCOM and TRADOC commands in an attempt to survey experts with a variety of experiences. While Officers had slightly more reliable ratings, there were no other appreciable NCO/Officer or FORSCOM/TRADOC differences. Thus, using an item-based method, either NCOs or Officers from FORSCOM or TRADOC could be used. Restricting the diversity of SMEs, however, raises the issue of standard acceptability. If the test and resulting standards were to be used at both FORSCOM and TRADOC sites, it would be prudent to survey SMEs from both commands.

Because both NCOs and Officers are affected by scores from job performance measures, it is advisable to use both in standard setting exercises. One could also argue that because airmen are affected by the standards, their judgments (i.e., incumbents' judgments) should be considered when defining those standards. The central issue here may be summarized by the question: Who are the users of the research results, and are they represented?

While it is important to survey SMEs from a variety of backgrounds, only SMEs who are directly familiar with airman performance are appropriate for examinee-based standard setting procedures. In most cases, NCOs work directly with airmen and consequently are more familiar with an individual airman's performance than are Officers. If an examinee-based method is used, it might be wise to use only NCO raters.

Number of judges. In addition to obtaining SMEs from diverse experiences, one must decide on the optimal number of judges. The optimal number of judges is determined to some extent by psychometric considerations, by the standard setting method employed, and by the number of qualified SMEs available. The number of judges is positively correlated with the reliability of the standard and negatively correlated with the amount of dispersion in the standard (Pulakos et al., 1989). Jaeger and Keller-McNulty (1986) suggest determining the necessary number of SMEs based on reductions of the standard error of the test standard and the standard error of measurement of the test. Cross, Impara, Frary, and Jaeger (1984) and Jaeger and Busch (1984) found that psychometric considerations are maximized with sample sizes of 20 to 30.

One must also consider the various types of raters being surveyed (e.g., NCOs and Officers from Site A and Site B). If main effects or interactions exist for rater type, a large number of raters is needed (e.g., 20 to 30 of each rater type). Thus, for the four types of raters suggested -- Site A NCOs, Site A Officers, Site B NCOs, and Site B Officers -- a total of 80 to 120 raters would be needed. Data from the U.S. Army's Synthetic Validity Project (Peterson et al., 1989), however, indicate that such a large number of raters is unnecessary. Furthermore, much standard setting research has been conducted with as few as five to eight raters (Andrew, & Hecht, 1976; Brennan, & Lockwood, 1980; Plake, & Melican, 1989; Skakun, & Kling, 1980).

The standard setting method often imposes practical constraints when determining the optimal number of SMEs. Methods implementing group discussions necessitate small- to medium-sized groups to prevent a few dominant SMEs from exerting too much control over other judges' decisions while still providing an adequate number of divergent opinions. Workshops with 20 participants are practical, but workshops involving more than 20 participants tend to be unmanageable.

Iterative process. As previously stated, the Jaeger method is the only standard setting procedure that prescribes an iterative process. All item-based methods, however, have been modified to include an iterative process. The primary purpose of the iterative process is to provide SMEs with an opportunity to reconsider their initial cutoff scores in light of potential consequences of those scores. The iterative process tends to follow one of several formats: (a) a presentation and individual consideration of normative data, (b) a presentation of the group's standard, (c) a group discussion allowing judges to debate the rationale underlying their cutoff scores, and (d) various combinations of (a), (b), and (c). The presentation of normative data does not require an iterative process. For example, normative data can be presented in the initial phase followed by an iterative process with a group discussion (Peterson et al., 1989). A group discussion, on the other hand, does necessitate an iterative process. The following discussion focuses on the group discussion as part of the iterative process.

A group discussion has been shown to reduce the variability in standards without significantly altering the standards (Jaeger & Busch, 1984; Norcini et al., 1987). By reducing the variability in standards, a group discussion thereby produces a more reliable standard. A few words of caution regarding the implementation of a group discussion iterative process, however, are in order. Pulakos et al. (1989) point out that individual judges' cutoffs, stated without justification, "can lead to a shift in judgment toward the central tendency of the group" (p. 29). To effectively evaluate differences in individual cutoff scores, the discussion must provide an opportunity to examine the rationale behind those cutoff scores. As in any discussion, a few dominant individuals are likely to unduly influence the group if not restrained. Therefore, a consensus discussion, the goal of which is to reach a general agreement among participants, is recommended rather than a convergent discussion, which requires unanimous agreement among participants. In addition, a skilled workshop leader is needed to maintain a controlled discussion.

In addition to a group discussion, the iterative process often includes the evaluation of normative data collected for the test in question. Some reviewers recommend that judges review normative data when setting performance standards (Hambleton, 1978; Shephard, 1980). By examining normative data, judges can evaluate the consequences of their recommended standard. Furthermore, the use of normative data has been shown to reduce the variability in standards (Cross et al., 1984; Jaeger, & Busch, 1984).

Hambleton and Powell (as cited in Pulakos et al., 1989) argue that the decision to use normative data should depend on the goals and constraints of the testing program. If the goal is to normally distribute examinees in terms of test scores, then emphasis on normative data is appropriate. However, if cutoff scores are to be used for selection purposes, too much emphasis on normative data is clearly inappropriate. A strong focus on normative data shifts the standard setting emphasis from "what performance should be" to "what performance is." In the present situation, cutoff scores are to be linked to selection standards. Therefore, emphasizing normative data would be a mistake; normative data should be used as a reality check only (i.e., to demonstrate the consequences of cutoff scores).

Training of judges. For the consideration of normative data to be effective, judges must be trained in their use. One cannot assume that judges can properly interpret even the "simplest" types of normative data (e.g., frequency distributions). SMEs must be taught how to properly read and interpret frequency distributions, graphs, etc. If more complex data are to be used (e.g., estimated item difficulty values), the meaning of these data must be carefully explained.

The second aspect of judge training involves insuring that SMEs fully understand the task they are to perform and familiarizing them with the test on which they will be setting standards. Explaining the standard setting procedure to be used may be fairly straightforward depending on the method being used. Clear, concise workshop instructions may be all that are necessary to ensure that judges understand the task at hand. On the other hand, Norcini et al. (1987) included a practice session in their explanation of the standard setting procedure being used. To familiarize SMEs with the test under consideration, Cross et al. (1984) and Jaeger and Busch (1984) had judges actually complete the test under approximately normal test administration conditions. At the very least, the instructions should include information about the way the test was administered and scored.

If an examinee-based method is used, training in the avoidance of halo error should also be conducted. As previously stated, examinee-based methods require the use of judges who are extremely familiar with the KSAs covered by the test as well as the performance of the individuals being classified in regard to those KSAs. Also, there is a positive relationship between rater-ratee familiarity and halo error (i.e., the more familiar the rater is with the ratee the more likely he or she is to commit halo error). Pulakos and Borman (1985) have developed a rater training program which has been shown to reduce rating error. Thus, some sort of rater training program should be conducted if an examinee-based method is used.

Number of standards. An additional issue concerns the number of standards desired. Is a single pass/fail score appropriate, or would several levels of performance standards be more beneficial? In many testing situations, several levels of performance are defined with performance below a certain point deemed unacceptable. In education for example, 90% correct or greater is often regarded as outstanding, 80% to 89% correct is superior, 70% to 79% correct is acceptable, and 69% correct or below is unacceptable. Although not explicitly stated, the purpose of various levels of performance is to encourage individuals to strive for improvement. Because the goal in war is to be better skilled than the enemy, airmen should never be encouraged to "rest on their laurels" once they have met the minimum performance standard. In maintaining job performance skills, the goal should be to strive for perfection. For these reasons, it is suggested that standards be established to differentiate among several levels of performance (e.g., unqualified, qualified, superior, and distinguished).

Performance standard definitions. A final consideration is the performance definition against which standards will be set. Performance definitions, in concept, determine what it means for an airman to be distinguished, superior, qualified, or unqualified. The question is whether the definition should be provided by researchers or by the SMEs. While no research could be found to demonstrate the superiority of either researcher- or rater-generated performance definitions, it seems prudent to begin the session with performance defined by the researcher. If the definition is completely out of line, raters can enhance it with the guidance of the researcher. If more than one workshop is to be conducted, the definition can be corrected at the first workshop. The corrected definition can then be used in subsequent workshops.

A Comparison of Guttman Scaling to Latent Trait Analysis

Guttman Scaling

The second standard setting method discussed by Guion (1978) is a priori scaling. This method entails the formation of a reference scale comprising a subset of test components (e.g., items, tasks, steps) that can be used to interpret total scores. One example of this type of scaling is Guttman scaling (e.g., Guttman, 1944).

A Guttman scale consists of test components ordered in such a way that an individual's correct response to (or endorsement of) a given component signifies his or her having supplied correct responses to (or endorsements of) all preceding components. For example, if a respondent correctly answers the fourth item from a test of items constituting a Guttman scale, then we know that the respondent has also correctly answered items 1, 2, and 3. If this same respondent did not correctly answer item five, then no other subsequent items would be answered correctly.

For example, the following set of items could be expected to constitute a Guttman scale:

1. We should continue to participate in the United Nations.
2. The United Nations is a constructive force in the world.

3. The United Nations is our best hope for peace.
4. The United Nations is the savior of all people.

Individuals positively endorsing item 4 would, in all likelihood, also endorse items 3, 2, and 1 positively, whereas individuals endorsing item 2 would probably not endorse items 3 or 4. A matrix for a Guttman scale of four individuals by the four items, containing 1's for correct responses and 0's for incorrect responses, produces a triangular pattern of responses:

	1	2	3	4
1	1	0	0	0
2	1	1	0	0
3	1	1	1	0
4	1	1	1	1

From a mathematical perspective, Guttman scaling is a deterministic model. That is, there is no error in the trace lines (curves relating standing on a trait to the probability of correctly responding to an item) of the items that make up the scale. At each point along the trait continuum (whether the trait be some ability, attitude, or other unidimensional characteristic), the probability of correctly responding to an item is either 0.0 or 1.0. Examples of these trace lines are provided in Figure 1. This figure reveals the origin of the term "step function" to describe these trace lines. The methods for developing a Guttman scale are termed "scalogram analysis" (e.g., Guttman, 1950).

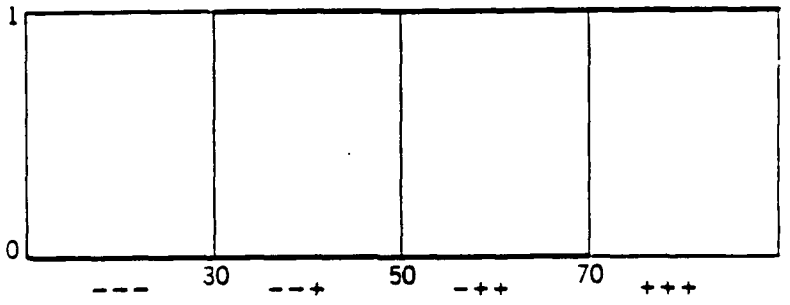


Figure 1. Three perfect monotone items.

The notion of a Guttman scale is very appealing because it provides an ordering similar to that given by many physical measurements (e.g., length). Nevertheless, application of Guttman scaling is highly unrealistic for most purposes; virtually no items fit the model. Aside from its impracticality, there are other criticisms of this approach, including its provision of only ordinal measurement (as opposed to interval or ratio) and the insufficiency of the triangular response pattern presented above to signify the existence of a Guttman scale. (The triangular pattern can be forced to appear by selecting items that vary greatly in difficulty. This process usually results in the violation of the requirement that the underlying attribute be unidimensional.)

Latent Trait Analysis

The third standard setting method discussed by Guion (1978) involves latent trait analysis. Latent trait theory is recognizable under many different names, including item response theory, item characteristics curve theory, and latent structure analysis. The use of the term "latent trait" does not suggest that traits are static. It refers instead to a mathematical model by which the relationship of item responses is tied to an underlying dimension (Guion & Ironson, 1983). For ease of reference and definition, we will use the term "item response theory" because we feel it best captures the meaning of the concept.

Item response theory (IRT) attempts to overcome some well-documented shortcomings in the construction, use, and evaluation of today's educational and psychological tests. One shortcoming is the dependence of item statistics (such as item difficulty and discrimination) on the examinee sample. Difficulty levels (p-values, defined as percent correct for an item) are higher when calculated on a sample of examinees that are of higher ability than those in the population from which they were drawn. Discrimination levels also tend to be higher as the sample of examinees becomes more heterogeneous. A second shortcoming of classical test theory lies in the manner in which examinees are compared. Ability estimates for examinees are usually limited to situations in which the same or parallel test items are administered. The difficulty of the items on a test and the ability levels of the examinees are confounded. Also, tests tend to be constructed for middle-ability persons and do not measure as well for high- or low-ability examinees. A third shortcoming of classical test theory is that it provides no performance estimations for individuals when confronted with test items. The probability that an individual will answer an item correctly is valuable information when adapting a test to an individual's ability level. There is also a presumption that the variance of errors of measurement is the same for all examinees. The capability of a test model to provide information as to the error of measurement on an individual level would help to provide a more precise estimate of the true ability of the examinee (Hambleton & Swaminathan, 1985).

Whereas Guttman scaling is deterministic, IRT is a probabilistic model. The trace lines, or item characteristic curves (ICCs), in IRT typically follow either a normal ogive or logistic function. This function, in turn, is typically defined by one, two, or three parameters, giving rise to one-, two-, and three-parameter models. These parameters are difficulty (defined as the point on the underlying attribute distribution, theta, that corresponds to a probability of .50), discrimination (a value proportional to the slope of the ICC at the theta value that represents the difficulty of the item), and pseudo-guessing (a value designed to account for guessing in multiple-choice tests that raises the lower asymptote of the ICC from zero to the probability of getting an item right by chance).

IRT can be used to define competency by providing a scale (theta) which represents competency and that is sample independent. One's standing on theta can be estimated to a desired degree of accuracy for each individual, using different items for each individual that are targeted to that individual's level of competency. The individual's responses to the items are used to better estimate his or her standing on the latent trait. Given a large set of tasks

representing the domain of tasks for a job, ICC's for these tasks could be computed and used to determine an individual's standing on theta (i.e., his or her level of competency). Because theta is a standard scale (i.e., it is not dependent upon the sample used to generate it), its use solves many problems associated with more traditional ways of scoring proficiency tests (e.g., atypical patterns of correct and incorrect performance on the components of the test).

Proposed Standard Setting Procedure for the AFHRL/JPM Project

The overall goal of the current project is to establish minimum performance standards for the Aerospace Ground Equipment Mechanic (454X1). There are four instruments available for establishing job performance standards: (a) job knowledge tests, (b) hands-on tests, (c) interview tests, and (d) rating scales. The rating scales are comprised of continuously scored items; the remaining measures are scored dichotomously. The job knowledge tests follow a written, multiple choice format, whereas the hands-on tests are performance or worksample tests scored GO/NO-GO. The interview tests are a type of performance test and were developed to assess tasks that are important to the AFS but are too expensive, too time-consuming, too dangerous, etc. to assess with a hands-on measure. The examinee talks through the procedures necessary to perform the particular task and may be prompted by the interviewer/scorer. The interview tests, like the hands-on tests, are scored GO/NO-GO.

Because the job performance measures were developed as part of a selection and classification study, it is reasonable to assume that they adequately cover the AFS under consideration (Lipscomb & Dickinson, 1987). Therefore, the next question is whether to set standards on each test (e.g., each hands-on test) or to set standards on dimensions of job performance (e.g., all tests covering mechanics). Regardless of the level at which standards are set (i.e., test vs. dimension), these standards must be aggregated to form a standard for job performance. The methodology for aggregating standards is beyond the scope of the present paper. Standard setting procedures are basically the same whether standards are set at the test or dimension level; therefore, the following discussion is limited to the identification of an appropriate procedure for establishing standards at the test level. Because hands-on test performance is a more accurate predictor of job performance than any of the other available measures, the discussion will focus on setting standards on the hands-on tests. In addition, an examinee-based technique based on the use of archival data as well as an item-based technique requiring the use of SMEs are proposed.

Regardless of the technique used to establish performance standards, an overriding goal of the project is to establish several levels of competency. In keeping with that goal, we decided to define five levels of proficiency. The five proficiency or competency levels correspond to the 5-point rating scale and definitions developed for the JPM Project. Those proficiency levels and their behavioral definitions are presented in Figure 2.

Always exceeds acceptable level of performance	Displays exceptional knowledge/skill to consistently complete assignments and tasks properly; requires little or no supervision; completes tasks in minimum time.
Frequently exceeds acceptable level of performance	Displays considerable knowledge and skill to complete assignments and tasks properly; performs effectively with little supervision; completes tasks more quickly than the average first-term airman.
Meets acceptable level of performance	Displays good knowledge/skill in most aspects of the job; able to properly complete the majority of tasks; requires supervision only on difficult tasks and assignments; completes work in the same time as other first-term airmen.
Occasionally meets acceptable level of performance	Occasionally displays adequate knowledge about how to complete tasks and assignments; quality of work is inconsistent; requires direct supervision on most tasks to ensure quality and accuracy; usually completes tasks within required time.
Never meets acceptable level of performance	Does not display knowledge and skill necessary to properly complete tasks and assignments; unable to perform without direct supervision; often fails to complete assignments; performs more slowly than other first-term airmen.

Figure 2. Proficiency levels for setting job performance standards.

Examinee-Based Archival Technique

Cantor (1989) demonstrated that examinee-based procedures could be used to set standards on archival data (i.e., without the use of SME judgments). Given budget constraints and time resources necessary to conduct standard setting workshops, a similar use of archival data seems practical for establishing performance standards in the present project. Specifically, rating data could be used to map cutoffs on the hands-on tests.

Several methods will be employed in the comparison of rating data from the hands-on tests. First, however, a description of the method by which the overall hands-on test score is computed is necessary. The Aerospace Ground Equipment (AGE) hands-on tests were administered in the work setting and consisted of tasks which were common across the specialty. The incumbents were instructed to perform the task according to technical order (TO) specifications, and were permitted to reference TO manuals or other written information as needed. The steps within those tasks were scored by a test administrator as either a GO (correct) or a NO-GO (incorrect). Each of the steps within the tasks had been assigned a weight by a senior Non-commissioned Officer (NCO) during scoring

workshops held prior to data collection. Weights were summed across all steps for a task to define the "base score" for that particular task. Weights for each step that the TA scored as a GO were summed, divided by the base score, and multiplied by 10. Each task score, equated on a 10-point scale, allows comparisons across tasks or for the computation of a composite task score by summing or averaging the individual task scores. For our analyses, we propose to use an average composite task score.

As described previously, the five levels of proficiency described in Figure 2 correspond to the 5-point rating scale used in the JPM Project. We will use two types of ratings collected as part of the JPM Project: Global Technical Proficiency (GTP) and Task Proficiency (TP). GTP refers to how skilled a person is at performing the technical aspects of the job, ignoring interpersonal factors (i.e., willingness to work, cooperation with others) or situational factors (i.e., lack of tools, parts, or equipment). The GTP rating is an evaluation of the quality of an individual's work across tasks. TP ratings refer to how skilled a person is at performing a specific task (i.e., the 15 hands-on tasks). By definition, these TP ratings exclude interpersonal and situational factors. TP ratings provided for each of the 15 hands-on tasks, are based on the question, "At what level of proficiency could this individual perform this particular task?" Thus the GTP rating is an overall rating of technical proficiency (i.e., across all technical aspects of the job), and the TP rating points to proficiency on one task in particular.

The GTP and TP ratings were obtained from four sources: (a) supervisor, (b) peer, (c) incumbent, and (d) test administrator (TA). For each examinee, there were up to three peer ratings. In our analyses, peer ratings will be averaged to yield a single peer rating. The GTP ratings were provided by supervisors, peer(s), and the incumbent. The GTP ratings summarize performance observed across time and tasks. The TA is a trained test administrator who does not interact daily with the examinee. For these reasons, the TA is not qualified to provide GTP ratings. The TP ratings are available from all four sources. The TP ratings provided by the TA are recorded immediately following observed performance of the hands-on test. The ratings provided by the other three sources are based on recall of on-the-job performance.

To provide a complete picture of performance, the two types of ratings (GTP, TP) obtained from the four sources (supervisors, peers, incumbents, and test administrators), will be combined in four different ways to categorize examinees into five proficiency levels. A mean GTP rating for each examinee will be computed by averaging the GTP supervisor (GTP-S), peer (GTP-P), and incumbent (GTP-I) ratings. Three TP ratings for each examinee will be calculated. The first, labeled TP-TA, is the TP rating supplied by the TA. The second, labeled TP-SPI, is an average rating based on the TP supervisor (TP-S), peer (TP-P), and incumbent (TP-I) ratings. The final TP rating, labeled TP-ALL, is the mean of the TP-S, TP-P, TP-I, and TP-TA ratings. These combinations are presented in Table 1.

Table 1

Methods of Categorizing Examinees into Proficiency Levels

Label	Type	Source	Combination Method
Mean GTP	GTP	S, P, I	Mean
TP-TA	TP	TA	Actual
TP-SPI	TP	S, P, I	Mean
TP-ALL	TP	S, P, I, TA	Mean

Examinees will first be categorized into the five proficiency levels presented in Figure 2 according to each of the four methods outlined in Table 1. The hands-on task scores for examinees falling within each of the five proficiency levels will be averaged. The midpoints between those five mean hands-on task scores will be the cut-off. For example, consider a hands-on task (HO_1). The cut-off score that defines "occasionally meets an acceptable level of proficiency" (P_2) will be the midpoint between the average score for examinees classified as "never meets an acceptable level of proficiency" and the average score for examinees classified as "occasionally meets an acceptable level of proficiency." The same procedure will be used to identify cut-offs for "meets an acceptable level of proficiency" (P_3), "frequently exceeds acceptable level of proficiency" (P_4), and "always exceeds an acceptable level of proficiency" (P_5).

For the four categorization methods presented in Table 1, four cut-off scores--defining the five levels of proficiency--will be identified for each of the 15 hands-on tasks and for the averaged composite of those tasks. This will allow comparisons of the various classification methods.

Item-Based SME Judgment Technique

In selecting an appropriate item-based method, each method was examined from practical and psychometric perspectives. The Nedelsky method was promptly eliminated for two reasons. The primary reason is that the procedure necessitates a multiple choice format, which prevents its use with the hands-on tests. Secondly, the standards produced tend to be lower and less reliable than those obtained via other methods. The time-consuming nature of the Ebel method precludes it from further consideration.

As originally proposed, all item-based standard setting procedures require that judges examine and make decisions at the item level. Given that there are 15 hands-on tests ranging in length from 7 to 30 items, the use of a method requiring judgments on each item is prohibitively time-consuming. Only the Angoff and Jaeger paradigms can be modified for use in setting standards at the test level. A comparison of the adapted Angoff questions and the adapted Jaeger questions demonstrates the greater flexibility of the Jaeger procedure.

The following four questions are adapted from the Angoff method:

1. What is the expected score for a group of airmen who *always exceed* an acceptable level of performance?

2. What is the expected score for a group of airmen who *frequently exceed* an acceptable level of performance?

3. What is the expected score for a group of airmen who *meet* an acceptable level of performance?

4. What is the expected score for a group of airmen who *occasionally meet* an acceptable level of performance?

The wording for the second, third, and fourth questions sounds incompatible. It is possible to explain that the score we are after is the lowest score for a particular category of individuals. This explanation, however, is not sufficiently different from the concept of the Jaeger derived questions:

1. What is the minimum score an airman could obtain to be considered to *always exceed* an acceptable level of performance?

2. What is the minimum score an airman could obtain to be considered to *frequently exceed* an acceptable level of performance?

3. What is the minimum score an airman could obtain to be considered to *meet* an acceptable level of performance?

4. What is the minimum score an airman could obtain to be considered to *occasionally meet* an acceptable level of performance?

For the reasons stated previously and to avoid possible confusion in trying to explain the concept underlying the Angoff questions, the Jaeger method is selected over all other standard setting procedures.

The original Jaeger method requires three iterations. In the initial phase, no normative data is presented. In the second phase, normative data is introduced; and in the third phase, additional normative data is presented. To reduce the number of iterations, normative data should be presented in the initial phase. Specifically, the percentage of examinees who received each test score should be presented in the initial phase. In addition to the initial phase data, the second phase should present the percentage of examinees who would have failed the test given the standard set by the group in the initial phase. SMEs would then complete the same standard setting procedure, and the test standard would be derived from this final iteration.

As originally proposed, the Jaeger method includes a consensus or Delphi discussion between iterations. During the discussion, judges' share the rationale for divergent cutoffs with the intent of reducing variability in standards. Compared to the original Jaeger method, the two iteration modification reduces the demand on judges' time. However, total time demands

may still be greater than what SMEs can reasonably spare at any one sitting. Therefore, we propose that the instructions and initial rating phase be completed in a workshop setting. Rationales for initial standards would be collected via a questionnaire during the initial phase, compiled by the researchers, and mailed along with the second phase materials to the SMEs. Judges would complete the second phase on their own time.

The use of normative data necessitates that standard setting activities--for at least the initial phase--occur in a workshop setting. It also dictates the presence of a skilled workshop leader. In addition to written instructions, the workshop leader must provide oral instructions on the standard setting procedure itself, as well as the proper interpretation and utilization of normative data.

NCOs and Officers representing the Aerospace Ground Equipment Mechanic specialty should serve as SMEs for the standard setting workshops. The use of incumbents is not advocated because they are not currently involved in setting performance standards. Incumbents generally are not perceived as being expert enough for decision making activities such as standard setting. For this reason, standards are not likely to be readily accepted. A total of 10 to 15 NCOs are needed to reach acceptable levels of reliability (Peterson et al., 1989). However, it may be politically prudent to include representatives from several groups of judges (e.g., Officers, training experts, operational experts). If such is the case, 10 to 15 representatives for each group of experts are needed.

Standard Setting Tradeoffs

In most cases, when cutoff scores are used to make selection or classification decisions, the cutoff does not reflect mastery of the task. Rather, it reflects a point at which those scoring above are able to perform to a degree of proficiency termed (by whichever means) as adequate. When the domain is narrow and a homogenous set of items is used, mastery and non-mastery become distinct and clearly evident in measurement. Yet as the domain becomes more broad, and more heterogenous items are required, the overlap between mastery and non-mastery increases, as do the errors in classification.

There are two different standpoints by which to view classification--from an individual standpoint or an organizational standpoint. From an individual standpoint, when a selection decision is made there are two outcomes which can result; either you are hired, or you are not. From an organizational standpoint there are four outcomes: (a) selection of a competent individual (true positive), (b) rejection of an incompetent individual (true negative), (c) selection of an incompetent individual (false positive), and (d) rejection of a competent individual (false negative). The advantages of the first two outcomes (i.e., true positive and true negative) from an organizational viewpoint are many. Correctly determining that a person is right for the job can be measured in the positive impact that person has on the organization (i.e., productivity). Correctly determining that a person is not right for the job, while it cannot be measured because that person is not present, assumes that the negative impact or the smaller degree of positive input would inhibit the organization. Making a correct decision is, of course, going to be beneficial to the organization. The decisions which have negative impact (i.e., false

positive and false negative) on the organization need to be recognized.

A false positive occurs when an incompetent individual obtains a score above the cut point, but actually does not possess the knowledge or expertise being measured. Possible reasons for the error in classification are: measurement error, bias, lucky guessing, cheating, or selective preparation for the exam (studied the right items). A false negative occurs when a competent individual, who has in fact mastered the task, fails to obtain a score above the cutting point. Possible reasons for this error in classification are: measurement error, bias, illness, unknown behavioral fluctuations, or complexity of instructions (Swezey, 1981).

Both types of classification error have the potential to be costly to the organization. In areas where subject matter mastery is critical, incorrectly classifying non-masters as masters (false positive) can be quite serious, particularly if the actions of an individual affects the performance of others. When an individual who is actually a task master is termed a non-master (false negative), costs to the organization are not clearly evident. While false negative errors do occur, cost estimates are difficult to quantify. Unless large amounts of resources were allocated to recruiting and/or processing of the individual, or an individual is sent for remedial training where it was not necessary, the false negative error cannot be recognized without tracking the rejected individual through either subsequent testing for other positions or through pre-training testing. Subsequent mastery performance could then be disguised as practiced performance or gained experience. The resources needed to identify false negative individuals are significant and become prohibitive to the extent that the costs of the false negative error, while they may be great to the individual, are low to the organization.

A consideration should be made, however, of the costs of false positives in reference to the setting of cutoff scores. If the costs (whether they be time, money, or production) are high then cut scores should be set high. In areas where successful completion of tasks is critical, a high cutoff will eliminate those that are fairly competent but not task masters. Of the four possible outcomes of selection and classification decisions (i.e., true positive, true negative, false positive, false negative), those regarding false positives seem to be the most important to the organization to quantify.

If it were possible to develop a selection procedure with perfect reliability and validity, the two types of errors (false positive and false negative) associated with selection decisions would vanish and the true score of individuals would emerge. But as with any test there is a band of uncertainty about the regression line of the test scores. The problem with false positive and false negative errors is that the two are compensatory. As one decreases the false positive error rate by raising the cut-point, the false negative error rate increases, and conversely, as you lower the cut-point to identify those false negatives as true masters, your false positive error rate will increase. With the use of cutoffs to determine mastery vs. non-mastery states, the methods by which those cutoffs are determined must be rigorous and are discussed in our Standard Setting Methods section.

Utility Approaches for Evaluating the Consequences of Standards

Decision theory is an approach to selection that recognizes the importance of outcomes or consequences of selection decisions to individuals and organizations. Utility methods are used to evaluate the consequences of selection decisions. Specifically, utility methods can be used to evaluate the accuracy of decisions made based on an operational cutoff score. Utility approaches typically assume that a single cutoff score has been established. However, the accuracy of several cutoff scores can be evaluated by inserting various cutoff score values into the utility formulas.

Cascio (1982) describes two approaches: (a) the proportion of total correct decisions and (b) the proportion of correct "accept" decisions. The proportion of total correct decisions (PC_{TOT}) equally weights incorrect rejections (i.e., false negatives) and incorrect acceptances (i.e., false positives). Most organizations are not interested in false negatives so the two errors usually are weighted differently. The proportion of correct "accept" decisions (PC_{ACC}) is used when the organization desires to maximize the proportion of individuals selected who will be successful. The formula considers only correct acceptances and false positives.

Most utility methods go beyond a simple evaluation of the proportion of correct and incorrect decisions resulting from the operationalization of a particular cutoff score. The more advanced methods were developed to evaluate the efficiency of various selection devices. These methods define the quality of selected individuals as: (a) the proportion of "successful" individuals in the selected group (Taylor-Russell), (b) the selected group's average standard score on the criterion (Naylor-Shine), or (c) the dollar payoff to the organization resulting from the use of a particular selection procedure (Brogden-Cronbach-Gleser).

Taylor-Russell. Taylor-Russell is one of the most well known utility models, perhaps due to its relative simplicity. It considers the validity of the selection device, the selection ratio (i.e., the ratio of the number of available job openings to the total number of available applicants), and the base rate (i.e., the percentage of applicants who would be "successful" without use of the selection device). It should be noted that the validity coefficient used in the Taylor-Russell model is the coefficient for the device used to select the current employees. The selection ratio is applied to this population. Thus, in addition to evaluating the effectiveness of a particular cutoff score, the method can be used to evaluate the utility of various selection devices or procedures.

The Taylor-Russell method makes several assumptions. First, it is assumed that individuals are selected for a specified course of action (e.g., employment, advanced training, etc.) and that this course of action cannot be modified. It ignores rejected individuals and categorizes accepted individuals into "successful" and "unsuccessful" groups.

Cascio (1982) points out that the Taylor-Russell method is most appropriately applied in the following three circumstances. The first case is applicable to most clerical or technician positions. For these and similar

jobs, differences in ability beyond that which is minimally required do not yield differences in benefit. The second situation occurs in military settings where selection and classification decisions are made by dividing individuals into two or more groups on the basis of predictor scores. In placement decisions, all individuals remain in the organization, but they are treated differently (i.e., assigned to various AFS). The final condition occurs when differences in output are believed to occur but these differences are currently not quantifiable (e.g., nursing care, counseling).

While it is beneficial to estimate utility by examining the magnitude of the increase in the proportion of successful applicants (i.e., the success ratio), it is even more beneficial to attach cost estimates to expected payoffs. Sands (1973) describes the Cost of Attaining Personnel Requirements (CAPER) model which is based on the Taylor-Russell method. The CAPER model was designed to estimate the total cost of recruiting, selecting, inducting, and training a sufficient number of individuals to meet a specified quota of successful individuals. The model provides estimates for:

1. Number of applicants who must be recruited in order to meet the quota
2. Number of erroneous acceptances
3. Number of erroneous rejections
4. Number of applicants who will be accepted
5. Total cost of employing the ordinary selection procedure to meet the quota
6. Total cost of employing the experimental selection procedure to meet the quota

Thus, the CAPER model can be used to demonstrate the costs of recruiting, selecting, inducting, and training individuals as well as the costs of erroneous decisions. Because costs are presented in terms of dollar costs and the number of individuals processed, the results can easily be communicated to decision makers and others affected by recruiting and selection procedures.

Schmidt, Hunter, McKenzie, and Muldrow (1979) point out two major disadvantages with the Taylor-Russell method. These are also disadvantages with the CAPER model because it is based on the Taylor-Russell method. The first disadvantage is the requirement to dichotomize job performance. This dichotomization results in a loss of information regarding the levels of performance. For example, the performance of all those in the "successful" group is assumed to be equal in value, whether they barely exceed the cutoff score or whether they perform well above the cutoff score. Also, the performance of individuals in the "unsuccessful" group is assumed to be equal.

A second disadvantage, noted by Schmidt et al. (1979) is that the procedure used to set the standard between successful and unsuccessful job performance is arbitrary. However, Block (1978) and Popham (1978) argue that although standard setting methods are somewhat arbitrary they are not as illogical as Schmidt et

al. would lead the reader to believe. Given the above discussion on standard setting procedures, it is assumed that a reasonably acceptable standard can be set using a well-researched procedure.

Naylor-Shine. The Naylor-Shine utility method builds on the Taylor-Russell procedure by assuming a linear relationship between validity and utility. In other words, for any given cutoff score, the higher the validity of the selection device, the greater the increase in average criterion score (i.e., job performance) for the selected group over that observed for the total group. The Naylor-Shine method is similar to the Taylor-Russell model in that the validity coefficient used is the coefficient for the device used to select the current employees. The Naylor-Shine method, however, does not require the dichotomization of criterion scores. Thus, a standard is not necessary to evaluate the utility of various selection procedures using the Naylor-Shine procedure.

The basic equation underlying the Naylor-Shine method considers: (a) the mean criterion score (in standard score units) of all individuals above the predictor cutoff score, (b) the validity coefficient, (c) the ordinate of the normal distribution at the predictor cutoff score (in standard score units), and (d) the selection ratio. Therefore, the method can be used to answer several questions.

1. Given a *specified* selection ratio, what will be the mean criterion score of those selected?

2. Given a *desired* selection ratio, what will be the mean criterion score of those selected?

3. Given a desired improvement in the mean criterion score of those selected, what selection ratio should be used?

4. Given a desired improvement in the mean criterion score of those selected, what predictor cutoff score should be used?

Compared to the Taylor-Russell method, the Naylor-Shine model appears to have more widespread applicability. Cascio (1982) provides several examples of circumstances in which the Naylor-Shine method is most applicable. One example is of particular interest in the present context because it directly applies to Air Force selection and classification. A primary objective for the Air Force is to most effectively match recruits to an AFS. To do this, job performance must be forecasted for each recruit and each AFS, which can be accomplished through the use of a regression equation. By expressing predicted criterion scores in standard score units, the Naylor-Shine method can be used to assess the expected increase in mean criterion performance as a function of variation in the selection ratio. This information can be used to make decisions regarding the time and money that should be spent in recruiting activities.

A major limitation of both the Taylor-Russell and Naylor-Shine models is that they do not formally integrate the concept of dollars gained or lost into the utility index. (Although an adaptation of the Taylor-Russell method to estimate the dollar value of selection decisions [Sands, 1973] has been noted).

The methods simply assume that larger differences in the proportion of successful employees (Taylor-Russell) or larger increases in mean criterion scores (Naylor-Shine) will yield a financial savings for the organization. The Brogden-Cronbach-Gleser utility model formally considers the dollar value of decision outcomes.

Brogden-Cronbach-Gleser. The assumptions underlying the Brogden-Cronbach-Gleser utility model are fairly straightforward and in some cases similar to those for the Taylor-Russell and Naylor-Shine models. Like the previously discussed models, the Brogden-Cronbach-Gleser method assumes that the validity coefficient used is the coefficient for the device used to select the current employees. The method accounts for the average cost of testing each applicant and allows the decision maker to estimate the dollar payoff of accepted individuals. It is assumed that test score and dollar payoff are linearly related. Because the cost of rejecting individuals may or may not be of interest to organizations, the method allows organizations to account for those decisions as they see fit. Although the method recognizes that performance (i.e., performance on the predictor or the criterion) is a continuous variable, it does require setting a predictor cutoff score in order to fill all vacant positions.

Cascio (1982) points out that the Brogden-Cronbach-Gleser method is potentially the most versatile utility model available; nevertheless, it has not received widespread attention. Part of the reason for this lack of attention is the difficulty of obtaining the cost information required, specifically the standard deviation of job performance in dollars. Estimating the dollar cost of job performance requires a complicated, time-consuming procedure of estimating the dollar value of job behaviors for each employee. Schmidt, Hunter, McKenzie, and Muldrow (1979) suggest an alternative method for estimating the dollar standard deviation of job performance. Specifically, they recommend asking supervisors to estimate the dollar value of the goods and services produced by employees at various job performance levels. Although computer programmer supervisors in the Schmidt et al. study were able to provide such estimates, it may be that some supervisors would be unable to do so. Based on observations of the standard setting processes used in the U.S. Army's Synthetic Validity Project (Fotouhi, 1989), it is unlikely that supervisors of technical AFS will be able to make such dollar estimates.

Utility Measurement Component of Project A. In an effort that stemmed from Project A, the Army's long term program to develop a complete personnel system for selecting and classifying all entry-level Army enlisted personnel, Sadacca, White, Campbell, DiFazio, and Schultz (1989) attempted to determine the relative utility to the Army of different levels of performance in entry-level military occupational specialties (MOS). As part of the utility measurement component of Project A, their purpose was to provide information that would aid decision-makers in maximizing the payoff to the Army of improved selection and classification procedures (Sadacca et al.).

Much of the previous work in improving selection and classification procedures is difficult to apply to the military setting for two reasons. The first is that compensation in the military sector is quite different than that of the civilian sector. Salaries differ by rank, not by MOS, thus relative worth of a MOS cannot be determined with this metric as different positions are in

organizations. The second reason previous work is difficult to apply to military settings is in the overall mission of the military versus organizations. The military's goal is to defend the United States against threats to our national security, an organization's main purpose is to maximize profits by providing a high quality good or service in the most efficient manner possible. While it is possible to look at defense spending, putting a monetary value on our nation's defense is not an appropriate metric for maximizing preparedness for catastrophic events (Sadacca et al., 1989).

Exploratory workshops were conducted to develop a procedure which could obtain performance utility values for various MOS. Within the workshops, Army field grade officers considered how to define different levels of performance, how to measure the value of the different levels of performance, and what the context for these utility values ought to be. Once the procedures had been established, the officers made judgments regarding the specific utility values for the five performance levels. The mean values were stable across both the different officer specialties and the two different scaling methods developed in the exploratory workshops. It was shown that the officers shared similar perceptions of the relative worth of various levels of performance of enlisted occupations.

Possibly the most striking and promising aspect of this research effort was the reliability of the perceptions of utility for various levels of performance. Differences in patterns of results across MOS reflect differences in the way in which soldier performance contributes to organizational productivity (Sadacca et al., 1989).

Summary. Evaluating the consequences of decisions made with the operational use of a cutoff score or scores produced by a traditional standard setting procedure is warranted, especially in the present context. There is increasing talk of the need to reduce defense spending. A frequently proposed spending reduction method is to cut the number of personnel in all branches of the Armed Services. As the total number of personnel decreases, the need to select and retain highly qualified individuals increases. Utility models can provide an estimate of the cost of recruiting, selecting, and retaining highly qualified applicants.

Because of its relative simplicity, the use of the CAPER model (Sands, 1973) is recommended. While it may be difficult to obtain cost estimates of employing various selection procedures, such estimates are more easily obtained than estimates of specific job behavior costs or of performance level costs.

References

Andrew, B. J., & Hecht, J. T. (1976). A preliminary investigation of two procedures for setting examination standards. Educational and Psychological Measurement, 36, 45-50.

Block, J. H. (1978). Standards and criteria: A response. Journal of Educational Measurement, 15, 291-295.

Borman, W. C. (1975). Effects of instructions to avoid halo error on reliability and validity of performance evaluation ratings. Journal of Applied Psychology, 60, 556-560.

Brennan, R. L., & Lockwood, R. E. (1980). A comparison of the Nedelsky and Angoff cutting score procedures using generalizability theory. Applied Psychological Measurement, 4, 219-240.

Buck, L. S. (1977). Guide to the setting of appropriate cutting scores for written tests: A summary of the concerns and procedures (Technical Memorandum 77-4). Washington, DC: Personnel Research and Development Center, United States Civil Service Commission.

Cantor, J. A. (1989). A validation of Ebel's method for performance standard setting through its application with comparison approaches to a selected criterion-referenced test. Educational and Psychological Measurement, 49, 709-721.

Cascio, W. F. (1982). Applied psychology in personnel management. (2nd Ed.). Reston, VA: Reston Publishing Company, Inc.

Cross, L. H., Impara, J. C., Frary, R. B., & Jaeger, R. M. (1984). A comparison of three methods for establishing minimum standards on the National Teacher Examinations. Journal of Educational Measurement, 21, 113-130.

Dillon, R. F., & Stevenson-Hicks, R. (1983). Competence vs. performance and recent approaches to cognitive assessment. Psychology in the Schools, 20(4), 142-145.

Glaser, R., Lesgold, A., & Gott, S. (1986). Implications of cognitive psychology for measuring job performance. Paper prepared for the National Academy of Sciences.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. American Psychologist, 18, pp. 519-521.

Glass, G. V. (1978a). Standards and criteria. Journal of Educational Measurement, 15, 237-260.

Glass, G. V. (1978b). Minimum competence and incompetence in Florida. Phi Delta Kappan, 59, 602-605.

Green, B. F., & Wigdor, A. K. (eds.), (1988). Measuring job competency. Washington, DC: National Academy Press.

Guion, R. M. (1978). Principles of work sample testing III. Construction and evaluation of work sample tests (TR-79-A10). Alexandria, VA.: U.S. Army Research Institute for the Behavioral and Social Sciences.

Guion, R. M., & Ironson, G. H. (1983). Latent trait theory for organizational research. Organizational Behavior and Human Performance, 31, 54-87.

Guttman, L. (1944) A basis for scaling qualitative data. American Sociological Review, 9, 139-150.

Guttman, L. (1950) Chapters 2, 3, 6, 8, and 9 in Stouffer, et al. Measurement and Prediction. Princeton, NJ: Princeton University Press.

Halpin, G., & Halpin, G. (1987). An analysis of the reliability and validity of procedures for setting minimum competency standards. Educational and Psychological Measurement, 47, 977-983.

Halpin, G., Sigmon, G., & Halpin, G. (1983). Minimum competency standards set by three divergent groups of raters using three judgmental procedures: Implications for validity. Educational and Psychological Measurement, 43, 185-196.

Hambleton, R. K. (1980). Test score validity and standard setting methods. In R. A. Burk (Ed.) Criterion referenced measurement: The state of the art. Baltimore, MD: Johns Hopkins Press.

Hambleton, R. K. (1978). On the use of cutoff scores with criterion-referenced tests in instructional settings. Journal of Educational Measurement, 15, 277-290.

Hambleton. R. K., & Swaminathan, H. (1985). Item response theory: Principles and applications. Boston, MA: Kluwer-Nijhoff Publishing.

Hambleton, R. K., Swaminathan, H., Algina, J., & Coulson, D.B. (1978). Criterion-referenced testing and measurement: A review of technical issues and developments. Review of Educational Research, 48(1), 1-47.

Jaeger, R. M. (1976). Measurement consequences of selected standard-setting models. Florida Journal of Educational Research, 18, 22-27.

Jaeger, R. M., & Busch, J. C. (1984). The effects of a delphi modification of the Angoff-Jaeger standard-setting procedure on standards recommended for the National Teacher Examinations. Paper presented at the joint annual meeting of the American Educational Research Association and the National Council on Measurement in Education, New Orleans, LA. (ERIC Document 246 091).

Jaeger, R. M., & Keller-McNulty, S. (1986, July). Procedures for eliciting and using judgments of the value of observed behaviors on military job performance tests. Prepared for the Committee on the Performance of Military Personnel, Commission on Behavioral and Social Sciences and Education, National Research Council/National Academy of Sciences.

Koffler, S. L. (1980). A comparison of approaches for setting proficiency standards. Journal of Educational Measurement, 17, 167-178.

Latham, G. P., Wexley, K. N., & Pursell, E. D. (1975). Training managers to minimize rating errors in the observation of behavior. Journal of Applied Psychology, 60, 550-555.

Lipscomb, M. S., & Dickinson, T. L. (1987). Test content selection. In H. G. Baker & G. J. Laabs (Eds.). Proceedings of Department of Defense/Educational Testing Service conference on job performance measurement technologies. San Diego, CA: Navy Personnel Research and Development Center.

Livingston, S. A., & Kastrinos, W. (1982). A study of the reliability of Nedelsky's method for choosing a passing score (Report No. ETS-RR-82-6). Princeton, NJ: Educational Testing Service. (ERIC Document NO. ED 218 361).

Mills, C. N. (1983). A comparison of three methods of establishing cutoff scores on criterion-referenced tests. Journal of Educational Measurement, 20, 283-292.

Norcini, J. J., Lipner, R. S., Langdon, L. O., & Strecker, C. A. (1987). A comparison of three variations on a standard-setting method. Journal of Educational Measurement, 24, 56-64.

Peterson, N. G., Owens-Kurtz, C., Hoffman, R. G., Arabian, J. M., & Whetzel, D. L. (1989). Army synthetic validation project: Report of Phase II results (Volume I) (ARI Technical Report). Alexandria, VA: U. S. Army Research Institute for the Behavioral and Social Sciences

Plake, B. S., & Melican, G. J. (1989). Effects of item content in initial judge consistency of expert judgments via the Nedelsky standard setting method. Educational and Psychological Measurement, 49, 45-51.

Poggio, J. P. (1984). Practical considerations when setting test standards: A look at the process used in Kansas. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Poggio, J. P., Glassnap, D. R., & Eros, D. S. (1981, April). An empirical investigation of the Angoff, Ebel, and Nedelsky standard-setting methods. Paper presented at the Annual Meeting of the American Educational Research Association, Los Angeles, CA.

Popham, W. J. (1978). As always, provocative. Journal of Educational Measurement, 15, 297-300.

Popham, W. J. (1975). Educational evaluation. Englewood Cliffs, NJ: Prentice Hall.

Popham, W. J., & Husek, T. R. (1969). Implications of Criterion-Referenced Measurement. Journal of Educational Measurement, 6(1), 1-9.

Pulakos, E. D., & Borman, W. C. (1985). Development and field test of the Army-wide rating scales and rater orientation and training program (ARI Technical Report 716). Alexandria, VA: U. S. Army Research Institute for the Behavioral and Social Sciences.

Pulakos, E., Wise, L., Arabian, J., Heon, S., & Delaplane, S. K. (1989). A review of procedures for setting job performance standards. Washington, DC: U.S. Army Research Institute for the Behavioral and Social Sciences.

Sadacca, R., White, L.A., Campbell, J.P., DiFazio, A.S., & Schultz, S.R. (1989). Assessing the utility of MOS performance levels in Army enlisted occupations. Washington, DC: U.S. Army Research Institute for the Behavioral and Social Sciences.

Sands, W. A. (1973). A method for evaluating alternative recruiting-selection strategies: The CAPER model. Journal of Applied Psychology, 57, 222-237.

Schmidt, F. L., Hunter, J. E., McKenzie, R. C., & Muldrow, T. W. (1979). Impact of valid selection procedures on work-force productivity. Journal of Applied Psychology, 64, 609-626.

Shepard, L. (1980). Standard setting issues and methods. Applied Psychological Measurement, 4, 447-467.

Skakun, E. N., & Kling, S. (1980). Comparability of methods for setting standards. Journal of Educational Measurement, 17, 229-235.

Swezey, R. W. (1981). Individual Performance Assessment: An Approach to Criterion-Referenced Test Development. Reston, VA: Reston Publishing Co.

Wigdor, A. K., & Green, B. F. (eds.), (1986). Assessing the performance of enlisted personnel: Evaluation of a joint-service research project. Washington, DC: National Academy Press.

Wood, R., & Power, C. (1987). Aspects of the competence-performance distinction: educational, psychological and measurement issues. Journal of Curriculum Studies, 19(5), 409-424.

APPENDIX A

Table 2

Cross-Reference Matrix of Studies

	Issues/ Critique	Definitions	Procedure Description	Procedure Comparison	Education/ Certification	Job Perf Standards	Application	Utility Model
Andrew & Hecht (1976)				X	X		X	
Block (1978)	X	X						
Brennan & Lockwood (1980)				X			X	
Buck (1977)	X	X	X					
Cantor (1989)				X		X	X	
Cascio (1982)	X	X	X					X
Cascio, Alexander, & Barrett (1988)	X							
Cascio & Silbey (1979)								X
Cross, Impara, Frary, & Jaeger (1984)				X	X		X	
Dillon & Stevenson-Hicks (1983)	X	X						
Emrick (1971)								X
Glaser, Lesgold, & Gott (1986)	X	X					X	
Glaser (1963)	X	X		X				
Glass (1978a)	X	X	X					
Glass (1978b)	X	X	X					
Green & Wigdor (1988)	X			X		X		
Guion (1978)		X	X				X	X
Guion & Ironson (1983)	X	X	X					
Halpin & Halpin (1987)				X	X			

Table 1 (continued)

	Issues/ Critique	Definitions	Procedure Description	Procedure Comparison	Education/ Certification	Job Perf Standards	Application	Utility Model
Halpin, Sigmon, & Halpin (1983)			X	X	X			
Hambleton (1978)		X			X			
Hambleton (1980)	X	X	X					
Hambelton, Swaminathan, Algina, & Coulson (1978)		X	X	X				
Jaeger (1976)	X	X						
Jaeger & Busch (1984)				X	X		X	
Jaeger & Keller-McNulty (1986)	X	X	X	X		X		
Karni & Lofsness (1985)					X		X	
Koffler (1980)				X	X		X	
Mills (1983)				X	X			
Norcini, Lipner, Langdon, & Strecker (1987)					X			
Plake & Melican (1989)	X				X			
Poggio (1984)				X	X		X	
Poggio, Glasnapp, & Eros (1981)				X	X		X	
Popham (1978)	X	X						
Popham & Husek (1969)		X	X	X				
Pulakos, Wise, Arabian, Heon, & Delaplane (1989)	X	X	X	X		X		
Sadacca, White, Campbell, Difazio, & Schultz			X			X	X	X
Sands (1973)								X

Table 1 (continued)

	Issues/ Critique	Definitions	Procedure Description	Procedure Comparison	Education/ Certification	Job Perf Standards	Application	Utility Model
Shepard (1980)	X		X					
Shikiar & Saari (1985)	X					X		
Skakun & Kling (1980)				X	X			
Wigdor & Green (1986)	X					X		
Wood & Power (1987)	X	X		X				

ANNOTATED BIBLIOGRAPHY

Andrew, B. J., & Hecht, J. T. (1976). A preliminary investigation of two procedures for setting examination standards. Educational and Psychological Measurement, 36, 45-50.

Hypothesis/Goal

1. "Will different standard setting procedures based upon similar assumptions [Ebel and Nedelsky] yield similar examination standards for comparable samples of test items?" (p. 46)
2. "Will different groups of judges using the same standard setting procedure in relation to the same examination content set different examination standards?" (p. 46)
3. "Will the average of judgments concerning examination standards made by individuals within each group differ from the consensus judgments of the group as a whole?" (p. 46)

Participants

Two groups of judges (n = 4 per group, total n = 8) drawn from test committees that contributed items to nationally administered certifying exam in the health professions.

Method

Test consists of 180 multiple choice items. Test was split using odd-even method of assigning items to subtest. Ebel method used with even numbered items; Nedelsky with odd.

After providing individual standards, group discussed each item according to Ebel or Nedelsky method, as appropriate, to arrive at consensus judgment.

Results/Conclusions

1. Ebel yields significantly different standards for comparable samples of test content.
2. Different groups of judges using same standard setting procedure in relation to same examination content do set similar overall examination standards.
3. Averages of judgments concerning examination standards made by individuals within each group do not differ significantly from consensus judgments of group as a whole.

Comment

Often cited study. Showed that individual standards don't differ from the group's standard as a whole.

Block, J. H. (1978). Standards and criteria: A response. Journal of Educational Measurement, 15, 291-295.

Rejoinder to Glass, G. V. (1978). Standards and criteria. Journal of Educational Measurement, 15, 237-260.

Glass argues all current methods of setting cutoffs on criterion-referenced tests are arbitrary. Block argues that even earliest standard setting procedures were based on some logic. More recent methods attempt to improve on logic of earlier procedures. While not perfect, recent procedures produce more defensible solutions than earlier methods.

Glass argues that because there are no nonarbitrary procedures, should abandon use of cutoffs in education and search for other solutions. Block argues that one advantage of using cutoffs is that it helps bring local school personnel, parents, and students into decision making process. Cutoffs also have positive impact on student learning (i.e., students who learn to mastery, better remember what they learned). Cutoffs on each objective to be learned serve as diagnostic tools to identify the modules that need to be repeated and those that don't.

Glass argues that one solution to standards problem is change - whether "rate of performance goes up or down." "Must still make some absolute value judgments as to how much change is enough, just as standard setters have tried to make such judgments as to what cut-off score is enough" (p. 294). This doesn't really solve the problem, just changes the form of the problem. Statisticians point out there are real problems with using change to measure performance anyway.

Brennan, R. L., & Lockwood, R. E. (1980). A comparison of the Nedelsky and Angoff cutting score procedures using generalizability theory. Applied Psychological Measurement, 4, 219-240.

Hypothesis/Goal

1. Consider generalizability theory as approach for characterizing and quantifying magnitude of error variances attributable to disagreement in rater judgments.
2. Illustrate this approach with experimental data.
3. Compare Angoff and Nedelsky procedures.
4. Examine impact of rater disagreement on some issues relating to reliability of measurement.

Participants

5 raters in health field set standards using Nedelsky and Angoff procedures on 126-item test. Same raters, same test, two standard setting procedures.

Results

1. Nedelsky mean cutoff was lower than Angoff mean cutoff.
2. Standard deviation of Nedelsky cutoff approximately twice as large as standard deviation of Angoff cutoff.
3. More variability attributable to differences in procedure means than to differences in rater means.
4. Reconciliation process using 5 raters and Angoff procedure resulted in 2 raters dominating discussion and stricter cutoff.

Conclusions

1. Differences in cutoffs may be due to:
   a. Procedure differences in way probabilities are assigned. Probabilities directly elicited in Angoff but are inferred by eliminating distractors in Nedelsky.
   b. Differences in ways minimum competency is conceptualized. Angoff allows raters to think of a single minimally competent person or a group of minimally competent people. Nedelsky allows raters to think only of a single minimally competent person.

Comment

The only study that uses the same group of experts to set standards using two different procedures (Nedelsky and Angoff). It's frequently cited by other standard setting authors.

Buck, L. S. (1977). Guide to the setting of appropriate cutting scores for written tests: A summary of the concerns and procedures (Technical Memorandum 77-4). Washington, DC: Personnel Research and Development Center, United States Civil Service Commission.

Setting cut scores is somewhat subjective. The desire is to set a score that will maximize the probability that those selected for the available jobs will be the more competent applicants, so that false-positives and false-negatives will be minimized, that an appropriate number of candidates will be provided, etc. Setting cut scores involves some value judgment as to minimal competency. It depends on the number of job openings and the number of available applicants.

Cut scores based on minimal competency should remain stable regardless of the number of job openings and the number of applicants. Cut scores should not be selected arbitrarily nor should a certain percentage of correct responses (e.g. 70%) be selected because that percentage has been used in the past.

Domain-referenced criterion-referenced test (DRT) - may have standard & cut score which do not coincide. 100% mastery of domain may be desirable but may be unreasonable to expect that. 90 or 95% may be acceptable as representative of mastery. Also, consider measurement error. Test may not measure all possible items in domain, examinee's observed score may not coincide with his/her true score.

Cantor, J. A. (1989). A validation of Ebel's method for performance standard setting through its application with comparison approaches to a selected criterion-referenced test. Educational and Psychological Measurement, 49, 709-721.

Hypothesis/Goal

Validate and examine reliability of Ebel method for setting minimum performance standards of Criterion-Referenced Systems Achievement Tests (C-R SATs) for Naval Enlisted Classification (NEC) in Strategic Weapon System (SWS), i.e., nuclear submarine technicians. C-R SATs used to establish job competency and readiness and to identify technicians that need re-training.

Compare Ebel to Contrasting Groups and Borderline Group methods.

Participants

Two levels of TRIDENT Technicians - watchstander and supervisor.

Method

Used C-R SAT data from Personnel Data Files which had been collected over several patrols.

Contrasting Groups - Study #1 - Classified masters and non-masters based on minimal competent profiles of watchstanders and supervisors (see p. 715). Once classified, examined C-R SAT data and plotted distributions for each group. Using original Ebel cutoff of 70%, categorized examinees as: master/pass, master/fail, non-master/pass, and non-master/fail.

Contrasting Groups - Study #2 (Borderline Group) - Comprised of supervisors only. Used different set of external criteria to classify masters and non-masters (i.e., average of previously obtained normalized SAT scores). Used cutoff of 51. (Not clear how cutoff was derived).

Results

Contrasting Groups - Study #1

- For watchstanders, 100% of those classified as masters passed. Those in the non-master/pass category, had significantly less experience and training than those classified as masters.

- For supervisors, non-master/pass group resembles watchstander non-master pass (i.e., they're borderline). Profiles for non-master/fail group were similar to non-master/pass group (i.e., they're borderline).

Contrasting Groups - Study #2

- SAT score criteria redistributed some supervisor masters and non-masters. Reduced number of false positives and increased number of true negatives.

- Borderline Group method would set cutoff of 73.5, but that's very close to 70 cutoff of Contrasting Group and Ebel.

Conclusion

1. Ebel is effective technique for setting standards.
2. Borderline Group more effective than Contrasting Groups for validating cutoffs. Easier to use, more reliable with population the size of the fleet, and conceptually compatible with Ebel.
3. Cutoff should be 2-3 points higher for watchstanders and supervisors to reduce false positives and false negatives.

Comment

The only study to use archival data to set standards with examinee-based methods. However, it's not clear how the procedures used in the Contrasting Groups - Study #2 is the same as the Borderline Group procedure.

Cascio, W. F. (1982). Applied psychology in personnel management. (2nd Ed.). Reston, VA: Reston Publishing Company, Inc.

In personnel selection, a classical validity approach is typically used in which the primary emphasis is on measurement accuracy and predictive efficiency. Simple or multiple regression is the basic prediction model used in the classical validity approach. Multiple regression is compensatory and assumes that high scores on one predictor can offset low scores on another predictor. In some situations (e.g., pilot selection), this assumption must be rejected and other selection models (i.e., multiple cut-off or multiple hurdles) must be used.

Decision theory attempts to overcome some deficiencies in classical approach, i.e., recognizes outcomes of prediction are of primary importance to individuals and organizations. Measurement and prediction (central themes to classical approach) are simply technical components of a system designed to make decisions about the assignment of individuals to jobs or treatments.

Unit Weighting - weight all predictors by 1.0. Appropriate when populations change from time to time or when predictors are combined into composite to boost effect size (and therefore statistical power). Does just as well as optimal weighting when weights are applied to new sample (see pp. 207-208).

Moderator Variables - e.g., gender, age, race, education. When correlation between predictor and criterion varies as function of classification on third variable; phenomenon known as differential prediction. Third variable is moderator variable. Utility of moderator variables rarely assessed. Moderator research requires very large sample sizes in each group, thus moderator variable effects are rarely assessed. One approach is subgrouping or using multiple moderators to construct profiles of scores.

Suppressor Variables - Little or no direct relationship to criterion but high intercorrelations with one or more predictors. Utility of suppressor variables in prediction not yet demonstrated.

Multiple Regression Approach - Particular values of predictors will vary widely across individuals although statistical weightings of each predictor will remain constant. Therefore, it's possible for individuals with widely different configurations of predictor scores to obtain identical predicted criterion scores. Compensatory model - assumes high scores on one predictor can compensate for low scores on another. Individuals rank ordered according to predicted criterion scores.

Multiple Cut-off Approach - Used when proficiency on one predictor cannot compensate for deficiency on another, i.e., when minimal level of proficiency on one or more predictors is crucial for job success and when compensation is not allowed. Selection is made from applicants who meet or exceed cutoff on all predictors. Failure on any one predictor disqualifies applicant from consideration. Assumes curvilinearity in predictor-criterion relationships - increasing levels of ability do not necessarily make person better qualified (i.e., more is not necessarily better).

No satisfactory solution developed for setting optimal cutoff scores in multiple cutoff model. In simple cutoff system (1 predictor) expectancy chart approach or Thorndike's "predicted yield" policy is used. Thorndike's "predicted yield" policy sets cutoff based on number of positions available during some future time period (e.g., 6 months), number of applicants to be expected during that time, and expected distribution of their predictor scores (based on local norms). Example, firm needs 50 secretaries in next year and anticipates about 250 applicants, selection ratio (50/250) is .20, thus, about 80% of the applicants will be rejected. Scores at 80th percentile on local norms plus or minus 1 standard error of measurement should suffice as acceptable cutoff. More than one predictor process becomes one of trial and error in which cutoffs for each predictor are set. For each pair of cutoffs, must determine how high average composite criterion score is for those selected compared to other possible cutoff score combinations. With more than 2 - 3 predictors, procedure becomes extremely tenuous. Expectancy charts can be used to depict likelihood of successful criterion performance to be expected from any given level of predictor scores. Expectancy charts computed from raw data and need not be limited to one variable or composite variable case or to discontinuous predictors.

Multiple Hurdle Approach - Cutoffs on some predictor may be used to make investigatory decisions. Applicants provisionally accepted and assessed further to determine whether should be permanently accepted. Most appropriate when training is long, complex, and expensive. In complete double-stage strategy, set two cutoffs on Test A, C_1 and C_2. Applicants who score above C_1 are unconditionally accepted, those who fall below C_2 are terminally rejected. Those who fall between C_1 and C_2 are provisionally accepted with final decision made on basis of Tests A and B.

Utility depends on (a) validity of a selection measure, (b) selection ratio (ratio of number of available openings to total number of available applicants), and (c) base rate (proportion of persons judged successful using current selection procedures). Taylor-Russell tables illustrate interaction of these parameters on success ratio (proportion of selected applicants who are subsequently judged successful). Ideally, the lower the selection ratio (few openings, lots of applicants) the better for the organization.

Decision-Making Accuracy - Evaluate accuracy of decisions made from setting cutoffs. Two approaches:

(a) proportion of total decisions made that are correct

$$PC_{TOT} = \frac{A + C}{A + B + C + D}$$

where A = correct acceptances
B = incorrect rejections
C = correct rejections
D = incorrect acceptances

Equally weights incorrect rejections and acceptances. Usually organization isn't very concerned about incorrect rejections, and differential weighting of these categories usually occurs.

(b) Proportion of correct "accept" decisions

$$PC_{ACC} = \frac{A}{A + D}$$ where A = correct acceptances
D = incorrect acceptances

Used when goal is to maximize proportion of individuals selected who will be successful.

Decision theory approach criticized because: (a) measurement errors are not considered in setting cutoffs and (b) use of mutually exclusive groups rather than a continuum of scores reduces precision.

Utility - Degree to which use of selection device improves quality of individuals selected beyond what would have occurred had that device not been used. Quality may be defined as: (a) proportion of "successful" individuals in selected group, (b) selected group's average standard score on criterion, or (c) dollar payoff to organization resulting from use of particular selection procedure.

Taylor-Russell Utility Model - Validity coefficient based on present employees who were selected using methods other than new selection procedure. Selection ratio applied to these people. Assumes fixed treatment selection (i.e., individuals selected for one specified "treatment" or course of action which can't be modified), ignores rejected individuals, and classifies accepted individuals into "successful" and "unsuccessful". Goodness of predictor is reflected only in terms of success ratio. When validity is fixed, success ratio increases as selection ratio decreases. Success ratio tells us that more people were successful, but not how much more successful.

Naylor-Shine Utility Model - Assumes linear relationship between validity and utility, i.e., given any arbitrarily defined cutoff on selection measure, the higher the validity, the greater the increase in average criterion score for selected group over that observed for total group. Increase in average criterion score to be expected from use of selection measure with given validity and selection ratio. Assumes validity coefficient based on concurrent validity model. Doesn't require dichotomizing employees into "satisfactory" and "unsatisfactory."

Naylor-Shine tables can be used to answer: (a) given specified selection ratio, what will be average performance level of those selected, (b) given desired selection ratio, what will be mean criterion score of those selected, and (c) given desired improvement in average criterion score of those selected, what selection ratio and/or predictor cutoff should be used? Neither Naylor-Shine nor Taylor-Russell formally integrate the cost of selection or dollars gained or lost into utility index.

Brogden-Cronbach-Gleser Utility Model - Accounts for the cost of selection. For more comprehensive discussion see pp. 222-223.

Taylor-Russell most appropriate when: (a) ability differences beyond minimum necessary to perform job do not yield differences in benefit, (b) placement decisions where individuals divided into 2 or more groups based on predictor

scores. All individuals remain in organization but receive different treatment, (c) differences in output believed to occur but are presently unmeasurable.

Naylor-Shine most appropriate when differences in criterion performance can't be expressed in dollar terms, but can assume that function relating payoff (i.e., performance under some treatment) to predictor score is linear. Useful in selection and classification situations like military.

Brogden-Cronbach-Gleser potentially most versatile utility model available, but hasn't received widespread attention to date. Most appropriate when criterion performance can be expressed in dollar terms and where can assume linear relationship between criterion and predictor.

Recommendation

Good summary of utility models which can be used to evaluate costs of selection procedures and cutoffs. Good place to start in trying to understand Decision Theory and utility analysis.

Cascio, W. F., Alexander, R. A., & Barrett, G. V. (1988). Setting cutoff scores: Legal, psychometric, and professional issues and guidelines. Personnel Psychology, 41, 1-24.

Purpose is to critically analyze and integrate legal, psychometric, and professional literatures as they address cutoffs and to summarize what's known and not known about use and misuse of cutoffs.

Case Law

Dent v. West Virginia, 1889 - Difficult standards OK unless procedure not valid.

Board of Regents of the University of the State of New York v. Tomanio, 1980 Upheld denial of chiropractic licensure of person who marginally failed licensure exam 7 times, said "lines must be drawn somewhere."

Contreras v. City of Los Angeles, 1981 - If applicants are rank ordered, cutoff may be mere formality.

Ratliff v. City of Milwaukee, 1985 - If trainees allowed to retake test until pass, cutoff has very little practical meaning.

Personnel Administrator of Massachusetts v. Feeney, 1979 - Cutoffs very important especially in situations where veterans given preference above everyone else if have passing score.

Administrative and Constitutional Law - General principle that administrative body (group that sets standards) has been given legal authority to give tests and set cutoffs. It's not up to the courts to second-guess that body unless there's compelling reason to. When set cutoff, administrative body must show that there's some rational relationship between cutoff and purpose of examination. Courts have upheld cutoffs for bar exam even when it's known that passing rates among various states may range from a low of 45% to a high of 90%.

Title VII Case Law - Not a lot of consensus among courts regarding appropriate standard to use in evaluating suitability of established cutoffs. *Rogers v. International Paper Company* (1975) - Cutoff so high that 40% of incumbents failed. Court said there must be some rationale provided for that level of cutoff. *Washington v. Davis* (1976) - Employer can seek to upgrade workforce by setting higher cutoffs. No guidance as to how high cutoff can be. General notion that it depends on situation (e.g., abilities of current workforce compared to those of qualified applicants in relevant labor market).

Pegues v. Mississippi State Employment Services of the Mississippi Employment Security Commission (1980) - OK to set cutoff that would eliminate bottom third of present employees. Based on notion that only top 2/3rds are satisfactory.

Thomas v. City of Evanston (1985) - Rejected cutoff that would eliminate 16% of incumbents. Said there was no evidence that 16% of employees weren't performing satisfactorily.

Berkman v. City of New York (1982, 1983, 1987) - Courts don't like cutoffs that would eliminate all incumbents. In Berkman, only 15% of highly trained military sample and no incumbents could pass physical ability test.

Cutoffs and Adverse Impact - Some organizations believe the issue of adverse impact can be avoided if set cutoff is so low that virtually all applicants pass. Tends to destroy credibility of testing process.

Thomas v. City of Evanston (1985) - First cutoff results in adverse impact so lower cutoff on next administration. Court said can't use this to argue that first cutoff was inappropriate.

San Francisco Police Officers' Association v. City and County of San Francisco (1987) - Some courts won't allow changes in cutoff or weighting system once established merely to reduce adverse impact. If no firm decision on weighting or cutoff, organization might be able to attempt to reduce adverse impact by certain adjustments.

Connecticut v. Teal (1982) - Bottom-line concept - if there's adverse impact in first test of multiple hurdles process, that test must be shown to be valid even if there's no adverse impact on total selection process.

** No single, mechanical, quantitative standard setting approach that's accepted or required by courts.

Cutoff should be consistent with results of job analysis, permit selection of qualified candidates, and allow organization to meet affirmative action goals.

** Courts look unfavorably on situations where expert's recommendation is ignored and organization sets cutoff on its own without any clear rationale (*U.S. v. Georgia Power Company*, 1973; Berkman 1982, 1983, 1987).

Plaintiffs often argue that any cutoff should be validated. Most courts found no merit in this approach. Validity, reliability, and utility issues refer to distribution of scores, inferences drawn from scores, and accuracy of decisions made, not to validity of specific score.

Defendant may have to show that test is valid, but doesn't have to show that cutoff was logical or justified. Plaintiff must show whether cutoff is appropriate. (*Gillespie v. State of Wisconsin*, 1985)

** Growing pressure to use more complex procedures for setting cutoffs (*Cuesta v. New York State Office of Court Administration*, 1987).

Norm-Referenced Methods

1. "Method of Predictive Yield" - Projected personnel needs, past history of proportion of offers accepted, large-sample distribution of applicants' test scores used to set cutoff that will yield number of new hires needed.
2. Set cutoff based only on distribution of applicants' test scores (e.g., at mean or 1 standard deviation below mean).
3. Set cutoff at point that provides sufficient pool to meet hiring requirements and minimizes adverse impact.

Advantages are simplicity and minimization of subjective judgment. Disadvantage is that relationship between cutoff and job performance is

unknown. Not likely to be acceptable where objective is to identify minimum competency (e.g., licensure).

Content-Related Validity Settings

Judgmental methods (Nedelsky, Angoff, and Ebel) and Empirical methods (Contrasting Groups and Borderline Groups) discussed here.

Of judgmental methods, Angoff generally preferred because it's at least as reliable as other judgmental methods and requires substantially less rater time and effort. Most promising variations involve iteration of judgment process, providing feedback on summary ratings, or feedback of item difficulties from normative samples.

Research suggestions:

1. Determine accuracy of each rater's judgments by including in items to be rated a subset of items for which normative data are available.

Problem is that norms for such items must come from group that's defined in same way as definition given to judges (e.g., minimally or barely competent).

2. Perform analysis of raters similar to item analyses (i.e., rater-total correlations analogous to item-total correlations) and eliminate raters on same basis as items are eliminated.
3. Use more carefully standardized definitions of groups for Contrasting Groups method (e.g., top 25% and bottom 25%).

Criterion-Related Validity Settings

Two approaches:

1. Set cutoff then determine whether cutoff will produce sufficient number of new hires and evaluate acceptability of expected mean job performance or expected false positive rate of those hired.
2. Find minimum level of job performance of false positive rate then set cutoff at level that would satisfy those requirements.

Research Suggestions (continued):

4. Find minimum level of job performance of false positive rate then set cutoff at level that would satisfy those requirements.
5. If performance measures constructed using Behaviorally Anchored Rating Scales (BARS) or Behavioral Expectancy Scales (BES), could incorporate consideration of minimum performance directly into each phase of scale development.
6. Where performance criteria already available, supervisors and incumbents could define minimum acceptable performance level on each criterion.
7. Set cutoff directly from minimum acceptable job performance level as function of test validity. Regression of predictor scores onto criterion provides prediction equation of calculating least-squares estimated predictor score associated with lowest acceptable criterion score.
8. Extend logic of expectancy chart. For each test score, proportion of validation sample who are above that score and above minimum acceptable level of job performance. Also determine maximum acceptable false positive rate.

9. Base cutoff on utility analysis.

Standards for Educational and Psychological Testing (1985) do not specify how to make cutoff decisions. Specify only information that should be made available to interested parties (standard errors of measurement, rationale for cutoff, validity of cutoff).

Principles for the Validation and Use of Personnel Selection Procedures (1987) may be interpreted as suggesting use of top-down selection. From psychometric point of view, this makes sense.

Guidelines for Setting Cutoffs

1. There is no single "best" method of setting cutoffs for all situations.
2. Standard setting process should begin with job analysis that identifies relative levels of proficiency on critical KSAOs.
3. Validity and job relatedness of assessment procedure are crucial considerations.
4. How test is used (norm-referenced vs. criterion-referenced) affects selection and meaning of cutoff.
5. When possible should consider relation of predictor and criterion scores.
6. Cutoffs should be high enough to ensure that minimum performance standards are met.
7. Cutoffs should be consistent with normal expectations of acceptable proficiency within workforce.

Recommendation

Most of the article addresses legal issues involved in setting cutoffs. These issues are not likely to present themselves in the context of military selection. The general notions of establishing cutoffs from psychometric and professional standpoint have some merit.

Cascio, W. F., & Silbey, V. (1979). Utility of the assessment center as a selection device. Journal of Applied Psychology, 64, 107-118.

Hypothesis/Goal
Demonstrate usefulness of utility theory through use of illustrative examples. Specifically applied utility theory to assessment center selection procedure.

Method
Using Brogden-Cronbach-Gleser continuous variable utility model, systematically varied:

1. Validity
2. Cost of assessment
3. Validity of ordinary selection procedure
4. Selection ratio
5. Standard deviation of criterion in dollars
6. Number of assessment centers

Assessment center payoff compared to (a) ordinary selection procedure and (b) random selection.

See article for details of how costs were estimated.

Results/Conclusion
The larger the criterion standard deviation (individual differences in criterion performance are large and significant) the greater impact of the assessment center. This may hold true even for predictors with low validity. High validity predictors may be less useful if individual differences in payoff are small. Thus, validity must be compared to other parameters.

Assessment center cost played relatively minor role in determining payoffs.

In conducting this study, authors made many assumptions and estimates (e.g., recruiting and training costs, employee tenure, etc.). Thus, there is some error in the calculations. Must be very careful when making these estimates.

Recommendation
Warrants further consideration if we plan to look at the cost of cutoffs. Formulas could be applied to any selection device not just assessment centers.

Cross, L. H., Impara, J. C., Frary, R. B., & Jaeger, R. M. (1984). A comparison of three methods for establishing minimum standards on the National Teacher Examinations. Journal of Educational Measurement, 21, 113-129.

Hypothesis/Goal

Angoff method usually used to set standards on NTE, but results in high failure rates. Goal is to evaluate Angoff along with Jaeger and Nedelsky for setting standards on NTE. Nedelsky of particular interest because consistently results in lower standards than other methods.

Participants

30 faculty members from 7 teacher-training institutions (15 for math exam, 15 for elementary exam). Each had 2 years experience teaching undergraduate teacher education courses in respective field. Judges randomly assigned to 1 of 3 panels (Angoff, Jaeger, Nedelsky) with members from the same institution balanced across panels.

Method

Incorporated corrections for guessing. (See pp. 117-118 for complete discussion.)

Three phases to evaluate normative feedback:

Phase 1 - Judges set standards on odd-numbered items with no normative data.

Phase 2 - Judges set standards on even-numbered items but were also given information on percentage of examinees estimated to know answer to each question.

Phase 3 - Judges set standards on odd-numbered items again and were (a) told percentage of examinees estimated to correctly answer each item, (b) taught how to evaluate stringency of their standards by consulting cumulative frequency distribution of scores from pilot test, and (c) Jaeger judges were told standards set by other judges in Phase 1 so they could evaluate the stringency of their standards.

During Phases 1 and 2, judges rated relevance of each question: 1 = crucial, 2 = important, 3 = questionable, 4 = not relevant.

At end of Phase 1, judges estimated percentage of prospective teachers that would fail test given the standard set in Phase I (subjective-failure rates).

At end of Phase 3, judges described their confidence in standard setting procedure used and congruence between topics covered by the exam and those covered in their curriculum.

Results

Normative feedback tends to reduce mean standard and amount of dispersion in standard.

Phase 1 standards significantly different from Phase 2 and 3 ($p < .05$), but Phase 2 and 3 standards not significantly different.

Nedelsky produced lowest standards (29.41), Angoff next (45.37), Jaeger highest (60.77) all $p < .05$. As methods originally/traditionally proposed, appropriate comparisons are Jaeger Phase 3 standards with Nedelsky and Angoff Phase 1 standards. These comparisons showed no significant differences for math test. Elementary test showed Nedelsky (27.54) significantly and substantially lower than Angoff (56.68) and Jaeger (58.[illegible])

Reliability lowest for Nedelsky, highest for Angoff, Jaeger in middle.

Standards stricter for more relevant items.

Conclusion

Nedelsky will yield desired lower standards, but Nedelsky standards are least reliable.

Judges have difficulty in identifying distractors a minimally competent individual would recognize as incorrect. "None of the above" distractor presents problems because it's not recognized as incorrect unless correct answer is known. "Which of the following statements is incorrect" requires judges to identify distractors that would be identified as correct. In other words, Nedelsky method can be confusing.

Normative data enhances psychometric characteristics of judgments, especially for Angoff and Jaeger.

Selected Phase 2 Angoff method because:

1. Yields standards that were more realistic than Jaeger standards and comparable failure rates across examinations. Nedelsky less stringent but more unreliable.
2. Greatest shift in standards occurred during Phase 2, and Phase 2 Angoff standards had best psychometric properties.
3. Judges using Angoff had more confidence in their standards and in their knowledge estimates than judges using other methods.

Recommendation

Couples the use of normative data and standard setting procedures that don't dictate its use (Angoff and particularly Nedelsky). May be beneficial if we want to use normative data with a method that doesn't require it.

Dillon, R. F., & Stevenson-Hicks, R. (1983). Competence vs. Performance and Recent Approaches to Cognitive Assessment. Psychology in the Schools, 20(4), 142-145.

The importance of accurate cognitive assessment and its relationship to effective instructional programming are discussed. Traditional methods of test administration are less than optimally sensitive to the cognitive abilities and processes under investigation. A disparity between competence and performance can occur across a variety of populations due to task and situational demands as well as to motivational and personality factors. Recent approaches aimed at lessening the gap between competence and performance are discussed, along with their strengths and weaknesses.

Effective assessment has three primary functions:

1. To indicate the developmental level of student in order to provide appropriate instructional material.

2. To provide a baseline by which subsequent performance can be compared.

3. To determine specialized needs in integrating low performers into the instructional setting.

Competence is defined as the examinee's actual level of cognitive functioning, if performance impediments were removed or eliminated. Competence is operationalized as the level of performance obtained under elaborative procedures beyond the performance level obtained under standard conditions. An activation model of cognition is assumed by the testing-for-competence paradigm. In general terms, this means that treatment effects will vary with respect to the extent to which a given testing condition serves to orient the examinee towards task demands.

Recommendation

Further research is needed to gain additional evidence of the utility of the testing-for-competence paradigm.

Emrick, J. A. (1971). An evaluation model for mastery testing. Journal of Educational Measurement, 8, 321-326.

Individually Prescribed Instruction (IPI) uses criterion-referenced procedures to set 85% cutoff on all skill tests. Is this cutoff appropriate for all skills? May be that for some tests, 60% indicates mastery and for others 90% indicates mastery.

Propose skill-mastery test model in which item and student information are combined to yield probability statements representing skill-mastery status. Advantages: (a) few simple assumptions, (b) provides for empirical determination of item measurement error likelihoods, (c) cutoff is provided by algorithm based on test properties and cost-benefit analysis of decision errors.

Assumptions:

1. For educational objective to be completely mastered, all component skills must be mastered. Degree of mastery determined by proportion or number of component skills that are mastered.
2. Component skills mastery tests consist of test items that are highly homogeneous in terms of content, form, and difficulty. Thus, each item response provides unbiased estimate of examinee's mastery status of that skill.
3. Examinee is either master or nonmaster given that mastery is assumed to be all or none. Two types of measurement error:
   a. Type I or alpha - responses lead to mastery conclusion when true status is nonmastery.
   b. Type II or beta - responses lead to nonmastery conclusion when true status is mastery.

   In other words, responses which correspond to examinee's true status are valid (i.e., items masters answer correctly and nonmasters answer incorrectly). Responses that don't correspond to examinee's true status are measurement error (i.e., "lucky guesses" for nonmasters and "careless errors" for masters).
4. Measurement error in single skill test can be approximated by calculating average inter-item correlation of responses to parallel, homogeneous items. Average inter-item correlation provides unbiased estimate of squared correlation between examinee's true mastery state and his item response.
5. Due to measurement error, decision errors will accrue regarding classification of examinees as masters or nonmasters. Can minimize errors through cost-benefit analysis of evaluative process variables. Three classes of variables:
   a. Statistical (item reliability, test length) - for tests of given length, more reliable items yield fewer errors. For given item reliability, increasing test length by adding parallel items increases reliability.
   b. Curricular - for objectives that are peripherally related to next skill level, errors are less important than for objectives that are critical to mastering next skill level.
   c. Psychological costs - masters erroneously classed as nonmasters (false negatives) costs are boredom, lower motivation, etc. For

nonmasters erroneously classed as masters (false positives) costs are confusion, etc.

Because variables can't be quantified, must make relative decision error costs and relative item error probabilities. Thus, optimization formula (p. 324). To use formula must decide:

a. Which type of error (Type I or II) predominates (e.g., true-false test should yield relatively more Type I [classify nonmaster as master], recall items should yield more Type II [classify master as nonmaster]).
b. How serious false positives and false negatives are.
c. Optimal test length assuming length can be changed.

Author gives example for IPI Math Placement Test.

Comment

This could be helpful in estimating costs of false positives and false negatives. The mathematics required aren't terribly difficult.

Glaser, R., Lesgold, A., & Gott, S. (1986). Implications of cognitive psychology for measuring job performance. Paper prepared for the National Academy of Sciences.

Presents a cognitive account of the components of skill, specific measurement procedures employed, and considers which aspects of measurement in the Services can best use these approaches.

Components of Skill

1. The contents of technical skills: The procedure of which they are composed.

2. The context in which technical skills are exercised: The declarative knowledge needed to assure that skill is applied appropriately and with successful effect.

3. The mental models or intermediate representations that serve as an interface between procedural and declarative knowledge.

Methods for Cognitive Task Analysis Measurement

Procedure Ordering Tasks - Involves the examinee ordering of tasks stopping short of actual performance of target task. The cases in which ordering tasks could be used involves (a) the possibility that the steps in the procedure could be carried out in several different orders and (b) constraints on ordering that would not be regulated by feedback the subject receives in the course of actually carrying out the procedure.

Sorting Tasks - Basic theory underlying the approach is that concepts are defined in the mind by a set of characteristic features. The general method involves having subjects place in separate piles on a table top the various things being sorted. A record is made of which items end up in which piles.

Realistic Troubleshooting Tasks - Where substantial amounts of diagnosis or other problem solving is involved in the job domain tasks can be revealing where they provide controlled opportunities for the subjects to actually do the difficult parts of their jobs.

Connection Specific Tasks - The breaking down of complex tasks into smaller components and solving smaller tasks.

What-How-Why Tasks - Several basic kinds of knowledge about circuit components, tools, or other important job artifacts are measured. "What" knowledge is the identification of an object. "Why" knowledge tells what the object is used for. "How" knowledge determines how the object works.

This paper also discusses the areas of the military where cognitive techniques have promise.

Glaser, R. (1963). Instructional Technology and the Measurement of Learning Outcomes: Some Questions. American Psychologist, 18, pp. 519-521.

Measurement of subject matter proficiency is the concern of this paper. Achievement measurement is defined as the assessment of terminal or criterion behavior involving the determination of student performance in reference to specified standards. Criterion-referenced measures depend upon an absolute standard of quality, while norm-referenced measures depend upon a relative standard.

"Underlying the concept of achievement measurement is the notion of a continuum of knowledge acquisition ranging from no proficiency at all to perfect performance. An individual's achievement level falls at some point on the continuum as indicated by the behaviors he displays during testing. The degree to which his achievement resembles desired performance at any specified level is assessed by criterion-referenced measures of achievement or proficiency. The standard against which a student's performance is compared when measured in this manner is the behavior which defines each point along the achievement continuum. The term "criterion," when used in this way, does not necessarily refer to final end-of-course behavior. Criterion levels can be established at any point in instruction as to the adequacy of an individual's performance. The point is that the specific behaviors implied at each level of proficiency can be identified and used to describe the specific tasks a student must be capable of performing before he achieves one of these knowledge levels. It is in this sense that measures of proficiency can be criterion-referenced.

Along such a continuum of attainment, a student's score on a criterion-referenced measure provides explicit information as to what the individual can or cannot do. Criterion-referenced measures indicate the content of the behavioral repertory, and the correspondence between what an individual does and the underlying continuum of achievement. Measures which assess student achievement in terms of a criterion standard thus provide information as to the degree of competence attained by a particular student which is independent of reference to the performance of others" (p. 519).

Items most suitable for measuring individual differences in achievement are those which will differentiate among individuals all exposed to the same treatment variable, while items most suitable for distinguishing between groups are those which are most likely to indicate that a given amount or kind of some instructional treatment was effective. Samples of test items are drawn from a population of items indicating the content of performance; the particular item samples that are drawn, however, are those most useful for the purpose of the kind of measurement being carried out.

Glass, G. V. (1978a). Standards and criteria. Journal of Educational Measurement, 15, 237-260.

Standards or "criterion levels" can only be determined arbitrarily.

1. Ordinary usage of words "standards" and "criteria" - language of performance standards is pseudoquantification. Numbers are applied meaninglessly to a question not prepared for quantitative analysis. General vs. specific descriptions of questions, tasks, or activities don't make it easier to set standards.

2. Trace evolution of notion of performance standards in criterion-referenced testing movement - Glaser's early writings characterized criterion-referenced testing as assuming a "continuum of knowledge acquisition ranging from no proficiency at all to perfect performance" and "degree of competence attained . . . is what's assessed".
When discussing behavioral objectives, Mayer suggested establishing minimum performance standards.
Popham used Glaser's notion of the word "criterion" to mean "standard," which later writers extended to mean, in addition to "standard," "mastery level," "cut-off score," or "pass-fail mark."
Thus, "criterion" in criterion-referenced testing has become synonymous with "standard," which Glass argues is confusing.

3. Analyze and critique 6 methods of setting performance standards on criterion-referenced tests -

   a. Performance of Others, i.e., norm referencing - Mastery level established as median test score earned by certain type of people.

   b. Counting Backwards from 100% - Desired performance is 100%, but due to measurement error, examinee fatigue, etc. perfection is impossible. Must decide what score less than 100% is acceptable. Glass argues these judgments are very arbitrary.

   c. Bootstrapping on Other Criterion Scores - Cutoff is determined by articulating test with external designation of "mastery." (e.g., Identify candidates who passed test. Define this group as "competent" based on other means. Look at distribution of competent group's scores on test in question and establish cutoff for separating competent from incompetent. Two problems:
      i. Both measures must be correlated 1.0 or examinees labelled "competent" with one measure will be labeled "incompetent" with other measure. To avoid this, cutoff is drawn arbitrarily, sometimes decision-theory techniques used but decision still arbitrary.
      ii. How do you rationalize cutoff selected? Why not rank order examinees and select top down?

   d. Judging Minimal Competence - Study test, item, or exercise and declare what "minimally competent" person score (e.g., Nedelsky, Ebel, Angoff). (Described Nedelsky and Ebel in some detail.) Upsetting

that different methods don't produce similar standards and that different judges using same method set disparate standards. If goal of various methods is to establish minimum competency, standards produced should be similar regardless of the different procedures used.

e. Decision-Theoretic Approaches - Given that cutoff is selected, what are consequences of that cutoff. Assumes false positives and false negatives are equal, therefore is highly arbitrary. Because identification of cutoff is so judgmental, why bother with determining cost of erroneous decisions.

f. "Operations Research" Methods - based on operations research approach of maximizing a valued commodity by finding optimum point on mathematical curve or graph. Example, separate but randomly equivalent groups taught until achieve various levels of proficiency on criterion-referenced test. Measure all groups on external measure of valued outcomes, i.e., retention scale, "life success," etc. draw graph relating scores on criterion-referenced test score and valued outcome. Score on criterion-referenced test for which valued outcome score is maximized is cutoff. Works only if scale is monotonic, (i.e., if graph bends towards baseline). Otherwise, criterion score that maximizes valued outcome is 100%. One way around the monotonic scale problem is to introduce a second valued outcome that's inversely related to degree of mastery on a criterion-referenced test (e.g., boredom). Must chose among composite outcomes, and this choice is arbitrary. If beyond some point on criterion-referenced test there are no gains in valued outcome, can set cutoff here. Works if one encounters curves with abrupt bends or corners, but these situations not likely to occur psychometrically.

4. Suggest how standards problem as conventionally defined can be ignored - "With respect to setting criterion scores on criterion-referenced tests, nothing may be safer and better than an arbitrary something" (p. 258). Standards are not absolute, and cutoffs can't be set without regard for their consequences (e.g., labeling incompetent surgeons as competent, effects on supply and demand, etc.)

Comment

Provides an overview of the problems inherent in standard setting. Gives the reader an appreciation for the arbitrary nature of standard setting.

Glass, G. V. (1978b). Minimum competence and incompetence in Florida. Phi Delta Kappan, 59, 602-605.

Draws heavily from Glass, G. V. (1978). Standards and criteria. Journal of Educational Measurement, 15, 237-260.

Criticizes Florida's minimal competency testing program for setting arbitrary standards while arguing that all standard setting is arbitrary.

Stresses the importance of evaluating the effects standards have on examinees (e.g., 35% failed math and 10% failed reading), which means that unless those examinees pass the test within the next two trials, they will not receive a high school diploma.

Comment

Provides some appreciation of the subjectiveness of standard setting. Mostly it seems like an opportunity for Glass to vent his frustration.

Green, B. F., & Wigdor, A. K. (eds.), (1988). Measuring Job Competency. Washington, DC: National Academy Press.

The Recommendation to Measure Competency

The Job Performance Measurement/Enlistment Standards Project of the Armed Services was established to examine the feasibility of measuring job performance and to link enlistment standards to job performance. The Committee on the Performance of Military Personnel, which was established to provide technical oversight to the project, expects the project to demonstrate several methods of measuring job performance adequately. The process of linking entrance standards to job performance is a more complex task requiring nontraditional methods and an expanded sense of policy perspectives.

The committee feels strongly that if the Joint-Service Project is to effectively communicate information about the performance of enlisted personnel and the implications of changing standards--either internally to military policy makers or to Congress--then the scoring scale of the job performance tests needs to be given some sort of absolute meaning. Scores should, in other words, communicate some sense of how well a person can do the job or, perhaps, how much of the job a person can do well. In contrast, scores currently say something about an examinee's relative standing with reference to all other examinee's, which is useful for ranking applicants but is not very informative about how a person at any particular score level will perform a given job. Measures of job competency would need to be referenced to some external scale of job requirements, not to the performance of other job incumbents.

The term competency as used here denotes a way of interpreting scores on a performance scale. It follows that there are degrees of competency. Unfortunately, the term has sometimes been used to signify a simple dichotomy, separating the competent from the incompetent.

That is not our meaning, nor our intent. As we shall argue, a performance dichotomy is neither implied nor necessary. In selection systems, minimum standards or cutoffs are placed on entrance tests, not on performance measures--on the input, not the output. Setting a particular input standard will result in a consequent output distribution of job performance scores, some low, some intermediate, some high. Policy makers must decide if the resulting distribution of performance scores is acceptable. They would be better able to make informed judgments about what is acceptable and what is unacceptable if performance scores could be interpreted in terms of what the job incumbent who scores at each level is able to do.

Guion, R. M. (1978). Principles of Work Sample Testing III. Construction and Evaluation of Work Sample Tests (TR-79-A10). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Abstract

Work sample tests should be relevant to the job, objectively constructed and scored, reliable, and capable of being scored on a standardized content-referenced scale. Detailed steps in working from job analysis to establishing test specifications are presented for assuring job relevance. Methods are suggested for scoring by a priori scaling, or by latent trait analysis, to provide a standard, content-referenced scale for scoring. Job samples should be evaluated primarily in terms of relevance and of generalizability. Seven principles of work sample testing are offered to researchers. This report, the third of four, is written for psychologists and others interested in research testing.

Guion, R. M., & Ironson, G. H. (1983). Latent trait theory for organizational research. Organizational Behavior and Human Performance, 31, 54-87.

Abstract

Latent trait theory is a relatively new development in measurement theory; emphasis in its application has been placed mainly on the measurement of ability, but potential areas of application extend well beyond into measurement of job and organizational characteristics, measurement of bias and adverse impact in equal employment compliance, attitude measurement, and the measurement of performance. The theories and models grouped under latent trait theory are therefore presented, in simple, nonmathematical form, for consideration by industrial and organizational psychologists. The rationale stems from problems encountered in classical psychometric theory with its practical dependence on distributions of attributes in samples and its theoretical dependence on parallel forms, problems alleviated by the use of latent trait analyses. This article presents some basic concepts and some available computer programs. Some controversies and unresolved problems are examined from a practical perspective.

Halpin, G., & Halpin, G. (1987). An analysis of the reliability and validity of procedures for setting minimum competency standards. Educational and Psychological Measurement, 47, 977-983.

Hypothesis/Goal
Investigate reliability and validity of 10 standard setting procedures: percentile, chance/ideal mean, Ebel, Nedelsky, Angoff, practitioners, masters group, nonmasters group, borderline group, and contrasting groups.

Participants
172 undergraduate education students and 83 practicing teachers took Missouri College English Test.

Method
Arbitrarily Selected Percentile - most competent 67% passed and least competent 33% failed.

Chance/Ideal Mean - (a) averaging lowest score in student sample group and expected chance score, (b) averaging actual mean score for student group and ideal mean score, (c) minimum passing score is midpoint between two averages.

Ebel, Nedelsky, and Angoff - 10 judges (5 university professors with training and/or experience in English, 5 graduate students in English education, 5 public school English teachers). Judges given detailed instructions and copy of test with correct answers marked. Judges worked independently. Reference point was beginning teacher minimally competent in English. Average across judges was cutoff for each method, respectively.

Practitioner Standard - mean for 83 practicing teachers.

Masters, Nonmasters, Borderline, and Contrasting Groups - mean for group of teachers identified as nonmasters became nonmaster standard. Mean for group of teachers identified as masters became master standard. Mean for group of marginal teachers became borderline group standard. Intersection of frequency distributions of masters and nonmasters became contrasting groups standard.

External Criterion Standard Setting - In 30 min. supervised session, 172 undergraduate education students wrote essay on general topic. Three faculty members evaluated essays using the holistic method and 10-point scale adapted from Coffman (1970). Average interrater reliability = .88. Two other faculty members categorized essays as competent and incompetent. Third faculty member categorized essays on which other raters disagreed. Of 152 essays judged competent, average of holistic ratings computed for each rater. Sum of these averages was minimum standard for writing sample.

Results
Considering equivalence an indication of reliability, phi coefficients computed using pass-fail decisions for all possible two-method combinations. Ranged from .16 to 1.00 with median being .53.

Using phi formula suggested by Berk (1976), validity coefficients computed for each of 10 standards. Validity coefficients were measures of extent to which

predictor (Missouri test) classification accurately represented criterion classification. Ranged from .20 for nonmastery method to .40 for practitioners.

Reliability and validity of item-judgment approaches - interjudge reliability using ANOVA, $\underline{r}$ = .84 for Ebel, $\underline{r}$ = .74 for Nedelsky, and $\underline{r}$ = .81 for Angoff. Pearson correlations between judges' ratings and item difficulty interpreted as indicators of validity. Range from .24 for Nedelsky and item difficulty, .47 for Ebel and item difficulty, and .57 for Angoff and item difficulty.

Conclusion

Borderline and practitioners more validly classify students judged competent and incompetent on external criterion.

High interrater reliability estimates for Ebel and Angoff consistent with findings from other studies. Greater validity for Ebel and Angoff over Nedelsky similar to Poggio et al. (1981) results.

Recommendation

All-encompassing comparison of standard setting procedure. "Validates" standard against external criterion, which is a relatively new approach to analyzing/evaluating standard setting methods. However, many researchers argue that a cutoff cannot be validated.

Halpin, G., Sigmon, G., & Halpin, G. (1983). Minimum competency standards set by three divergent groups of raters using three judgmental procedures: Implications for validity. Educational and Psychological Measurement, 43, 185-196.

Hypothesis/Goal

1. Do Ebel, Nedelsky, and Angoff methods produce different cutoffs for same standardized test?
2. Do same methods, used by raters with different competency levels, result in different standards for same test?
3. Are standards set by different methods the same across groups of raters (i.e., method by rater group interaction)?
4. What are implications for validity of results of this study?

Participants

3 groups of 5 raters (n = 15):

- Group 1 - Advanced doctoral students in English education. Mean of 3.5 yrs experience teaching English in public schools.
- Group 2 - High school English teachers. Mean of 4.6 yrs experience teaching high school English.
- Group 3 - University faculty with training and/or experience in English education. Mean of 10.8 yrs teaching experience at post-secondary level.

Method

Each judge rated all 90 items on Missouri College English Test, Form B using Ebel, Nedelsky, and Angoff procedures. Ebel-Nedelsky-Angoff order used (a) to minimize rating contamination due to carry-over effects and (b) represents logical progression in rating.

Raters worked independently. Rated each item using three methods before going on to next item.

Ebel - For each item, check 1 of 9 categories: Essential, important-easy, important-medium, acceptable-easy, acceptable-medium, acceptable-hard, questionable-easy, questionable-medium, questionable-hard. Reference point - beginning teacher minimally competent in English. Researchers provided percentage of examinees who should be able to answer questions in each category.

Nedelsky - Circle letter of options that beginning teacher minimally competent in English should be able to eliminate as incorrect. Correct option appeared in square.

Angoff - Lowest (20-25%) and highest (100%) probabilities appeared on rating form. Raters wrote in proportions differing from maximum and minimum.

Results

Reliability - .84 for Ebel, .74 for Nedelsky, and .81 for Angoff.
Estimated reliability for any of the groups - .62 for Ebel, .49 for Nedelsky, and .59 for Angoff.

No significant differences between groups across methods (i.e., mean cutoff across methods similar for three groups). Significant differences between methods across groups (i.e., mean cutoff different among methods applied within group). Significant interaction between groups and methods. (See Table 2, p. 191.)

Conclusion

1. Ebel, Nedelsky, and Angoff produce different standards on same test.
2. Different methods and different groups (groups differed in competency levels) set similar standards.
3. High school teachers set lowest standards with Nedelsky and highest standards with Angoff.
4. Ebel most stable. May be because raters' better understanding the task they are to perform. May be because task requires considering two dimensions rather than one.
5. Combination of methods across raters tended to stabilize cutoffs. Reinforces the use of several methods when setting standards.

Hambleton, R. K. (1978). On the use of cut-off scores with criterion-referenced tests in instructional settings. Journal of Educational Measurement, 15, 277-290.

Practical Suggestions for Setting Cut-Off Scores

1. Use several groups working together. Group size determined by importance of tests and number of domain specifications.

2. Introduce standard setting method. Short training session, including practice session. Discuss differences in standards.

3. Discuss domain specifications. Devote roughly equal amounts of time to each domain, with more complex or more important domains receiving more time.

4. Explain how tests will be used and the characteristics of individuals being tested.

5. Note any relationships among domains (e.g. mastery of one domain is prerequisite for performance in another domain).

6. Study differences in standards set by different groups.

7. Use past test performance data to modify cut-scores if necessary (e.g. if cut score is so high that, based on past data, a majority of people would have failed, then lower cut score accordingly).

8. Check percentage of "competent" and "noncompetent" as data become available. May need to change cut-off score, test items, etc. if percentages seem to be "out of line".

9. Review cut scores periodically. Priorities, politics, etc. may change.

Recommendation
Good guideline for developing, implementing, and evaluating a standard setting process.

Hambleton, R. K. (1980). Test score validity and standard setting methods. In R. A. Burk (Ed.) Criterion referenced measurement: The state of the art. Baltimore, MD: Johns Hopkins Press.

All standard setting methods involve judgment and are arbitrary.

How to select standard setting method: (a) importance of decisions, (b) amount of time available for standard setting, (c) resources (people and money) available for standard setting, (d) capabilities of judges (some methods require more knowledge of content and examinees to be tested than others), and (e) appropriateness of method for type of test being studied.

Judgmental Methods - Judges examine each item and indicate how minimally competent individual would perform.

Nedelsky Method - Applicable for multiple choice tests only. Judges identify distractors that minimally competent individual would identify as incorrect. Reciprocal of remaining options = minimum passing level (MPL) (e.g., 1 divided by 5 remaining options = .20). Sum MPL across all items = individual judge's standard. Average of individual judges' standards = test standard.

Ebel Method - Judges rate items along four levels of relevance (essential, important, acceptable, questionable) and three levels of difficulty (easy, medium, hard) in 3 x 4 grid. Judges assign items to one of twelve cells and assign percentage to each cell (percentage = percentage of items in cell that minimally competent individual would be able to answer correctly. Percentage is agreed upon by all judges.). Number of items in each cell x percentage for that cell. Sum all products divided by number of items = standard.

Angoff Method - Judges think of several minimally competent individuals instead of only one. Estimate proportion of minimally competent individuals who would answer each item correctly. Sum of probabilities = standard.

Jaeger Method - Through iterative process, judges from a variety of backgrounds using normative data ask (a) "Should every applicant be able to answer this item correctly? Yes or No?", and (b) if applicant does not answer this item correctly should he/she be denied employment? Yes or No?". Responses from groups of judges from same areas of expertise pooled and median is computed for each group. Minimum median across all groups = standard.

Empirical Methods - Livingston Method - Benefit and cost of decision linearly related to cut score.

Combination Methods - Based on combination of judgmental and empirical data.

Borderline-Group Method - (Zieky and Livingston) - Judges define minimally acceptable performance on content area being assessed. Submit list of individuals whose performances are so borderline that they can't be classified as acceptable or unacceptable. Administer test to these individuals. Median test score for group is standard or decision can be made to pass some other percentage of competent individuals.

Contrasting-Groups Method - (Zieky and Livingston) - Judges define minimally acceptable performance on content area being assessed. Identify masters and nonmasters. Administer test to these two groups. Plot score distributions of two groups, and intersection is taken as initial standard. Adjust standard up or down to reduce false positives (false masters) or false negatives (false nonmasters).

Contrasting-Groups Method - (Berk) - Administer test to instructed/trained and uninstructed/nontrained individuals. Set and evaluate standards based on percentage of false positives and false negatives. Berk offers several statistics to accomplish this. Works best in classroom situation.

Comment
Good summary of standard setting procedures.

Hambleton, R. K., Swaminathan, H., Algina, J., & Coulson, D.B. (1978). Criterion-referenced testing and measurement: A review of technical issues and developments. Review of Educational Research, 48(1), 1-47.

Purpose of paper is to show how criterion-referenced tests can have a wide variety of uses, and thier usefulness will be enhanced by technical developments that address their proper construction, validation, and interpretation. This paper is intended to serve as a review and integration of many recent developments in the field of criterion-referenced testing and measurement. The emphasis of the paper is on psychometric and statistical matters, however, that emphasis reflects the authors particular interests and competencies. Work needed in other related areas is mentioned in the concluding section.

This paper is organized into six setions: uses of criterion-referenced test scores, reliability of criterion-referenced test scores, determination of test length, determination of cut-off scores, test development and validation, and summary and suggestions for further research.

This paper serves as a companion paper to Hambleton's (1974) paper on testing and decision-making procedures within selected objectives-based instructional programs, and to provide an expanded discussion of one of the four major areas of use of criterion-referenced tests described in the monograph by Millman (1974). Content in this paper is centered on the use of criterion-referenced testing in instruction. Responses to Harris et al. (1974) measurement issues is also presented in this paper.

Jaeger, R. M. (1976). Measurement consequences of selected standard-setting models. Florida Journal of Educational Research, 18, 22-27.

Although standard setting models typically characterized as "judgmental" or "empirical," all require judgments of what is acceptable or unacceptable performance.

Judgmental Methods require judgments about domain behavior and sampled behavior. Test samples domain of interest (e.g., hands-on tests sample job performance domain). Use hands-on test scores to infer success or failure on job. Four threats to validity of inference:

a. Error among standard setters.
b. Error due to description of tasks in domain.
c. Inadequate sampling of domain (i.e., not enough items).
d. Inappropriate sampling of domain (i.e., wrong items).

Empirical Methods use behavior in related domain to set standards for domain of central interest (e.g., performance on core technical tasks used to infer performance on general soldiering tasks). Threats to validity:

a. Inadequate descriptions of either or both domains.
b. Unrepresentative sample of judges.
c. Inadequate number of judges.

Comment

Could review article in more detail. Given that so many of these issues are covered elsewhere, it's probably not very useful.

Jaeger, R. M., and Busch, J. C. (1984). The effects of a delphi modification of the Angoff-Jaeger standard-setting procedure on standards recommended for the National Teacher Examinations. Paper presented at the Joint Annual Meeting of the American Educational Research Association and the National Council on Measurement in Education, New Orleans, LA. [ERIC Document 246 091].

When groups consensus is required for standard setting judgments, dominant individuals unduly influence group judgments. Group's decision will be more extreme than average position of individuals within group. Standards tend to be higher.

Two advantages of discussion, without requiring group consensus, during standard setting procedures: (a) assuming random sample of participants, more precise estimation of mean without affecting mean recommended standard, (b) standard setters should be fully informed. Can presume that some members of standard setting group will have insights that others don't have. Discussing these insights will result in better-informed decision makers.

Depending on area of specialization, participants completed Reading or Social Studies subtest of NTE. Two standard setting opportunities followed. First, used Angoff method. Second, after instruction and practice session, were given actual test performance data and asked to reconsider original recommendations.

Two groups: Silence Group - Refrained from discussing recommendations before or during second judgment session. Discussion Group - Discussed initial recommendations before second judgment session with aid of group leader.

Second session judgments - Less variability than first session for silence and discussion groups. Even less variability for discussion group. Although variability reduced, mean standard remained stable.

** Sample sizes very small - N = 7 for Reading Silence, N = 7 for Reading Discussion, N = 6 for Social Studies Silence, N = 8 for Social Studies Discussion. Can't say much about rater types with these sample sizes. Reduction in variability is important finding.

Recommendation

Demonstrates influence of controlled discussion on variability in standards. Should be examined further if we decide to use a discussion when we set standards.

Jaeger, R. M., and Keller-McNulty, S. (1986). Procedures for eliciting and using judgments of the value of observed behaviors on military job performance tests. Paper presented to the Committee on the Performance of Military Personnel, Commission on Behavioral and Social Sciences and Education, National Research Council / National Academy of Sciences.

In setting standards for Project A tests, must set standards on clusters of tasks rather than on each task. In selecting tasks for testing, clustered tasks and tested a sample from cluster based on frequency and importance (e.g., tested 3 [?] first aid tasks, but 10 [?] first aid tasks performed on job.)

Problems:
Referent Population - Task specific or MOS specific (e.g., "Think about 100 soldiers who have just been admitted to 95B who are borderline in their knowledge of restraining a subject." vs. "Think about 100 soldiers who have just been admitted to 95B who are borderline in their knowledge needed to function satisfactorily as 95B.

Selection of Judges - TRADOC, FORSCOM, NCO, Officer, incumbents. Number of judges - 20-30 suggested (Cross et al., 1984; Jaeger & Bush, 1984).

Stimulus Material - Should be written and oral instructions. How to deal with guessing.

Training Judges - Explain testing conditions (e.g., outside in hot or cold, no study time, NCO scorers, round-robin procedure, etc.). Provide normative data. Delphi discussion.

Conventional Standard Setting Procedures as Applied to Performance Tests
Nedelsky Method - Won't work for dichotomously scored performance tests because it assumes multiple choice format. In civilian sectors, method often results in lenient standards compared to standards set with other procedures.

Angoff Method - Would work. For each task, raters indicate percent of minimally competent individuals who would perform each step correctly (e.g., "Think about 100 soldiers who have just been admitted to 95B who are borderline in their ability to restrain a suspect. What percentage would position suspect correctly when applying handcuffs?", etc. for each step in restraining suspect task.)

Ebel Method - Most tests have too few items to put in 3 (difficulty) x 4 (relevance) grid. Asking "What percentage of items should minimally competent soldier be able to answer correctly?" is same as "What should test standard be?" Not likely to be reliable. Could be applied to overall job performance test to yield standard for entire MOS. Problem - long tests would receive more weight than short tests due to method for calculating Ebel's method weights.

Jaeger Method - No referent population problems. May yield too strict standards. Ask "Should every enlistee who is accepted for this MOS be able to

perform this task? Yes or No". Tests mirror Soldier's Manual standards. If raters go "by the book" standards will probably be too high.

Borderline-Group or Contrasting-Group Methods - More effective with continuously scored items (e.g. time to fire weapon, accuracy). Classify people as "unacceptable", "marginal", or "acceptable" then collect data on at least one group.

Recommendation

Probably deserves more consideration because it speaks to standard setting in the Army's Project A, which has some similarities to JPM. Pulakos, E., Wise, L., Arabian, J., Heon, S., & Delaplane, S. K. (1989). A review of procedures for setting job performance standards. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences--is probably more comprehensive and clearer.

Karni, K. R., & Lofsness, K. G. (1985). Determination of passing scores on certification examinations: An unresolved issue. Journal of Allied Health, 14, 415-426.

Hypothesis/Goal
Examined results obtained from certification applicants and practitioners on national certification examination for clinical laboratory scientists (medical technologists) using modified Angoff procedure. Major question concerned appropriateness of cutoff score. Second was differences in examination performance between certification applicants and practitioners.

Participants
1,868 certification applicants. 111 practicing laboratory scientists selected by 40 volunteer laboratory managers. (Each manager (n = 51) selected 3 subordinates: one each of minimum, average, and maximum competence.)

Method
50 experts (10 from each of following 5 fields; clinical chemistry, hematology, immunohematology, microbiology, laboratory practice) determine cutoff using Angoff procedure. Final cutoff is average cutoff across judges subtract 4 standard errors of measurement to minimize false negatives.

Results
20.6% of the certification applicants failed. 79.3% of the minimally competent practitioners failed. (A total of 59.5% practitioners failed.) Mean of maximum competent practitioners was 13 points below mean of certification applicants.

Conclusion
Pass rate differences may be due to:

1. Motivation - applicants more motivated than practitioners.
2. Preparation Time - associated with motivation
3. Test-Taking Skills - applicants probably more test-wise than practitioners
4. Examination Content - exam probably reflected more current knowledge. practitioners may not keep up with advances
5. Age - based on another study, older people tend to score lower on certification exams.

Cutoff too high: "... examination investigated in this study is supposedly geared to practitioners of minimum competence. The fact that these persons scored poorly (only 20.6% passed) suggests that the examination is at a much higher level than that of minimum competence."

Current procedure (Angoff with 4 standard error of measurement downward adjustment) may be inappropriate. Judges may be setting standard too high to begin with.

Comment
Illustrates the importance of considering the consequences of decisions made by using cutoffs (i.e., number who pass and fail) when setting standards.

Koffler, S. L. (1980). A comparison of approaches for setting proficiency standards. Journal of Educational Measurement, 17, 167-178.

Hypothesis/Goal
Compare Contrasting Groups to Nedelsky.

Participants
Panelists (no more than 10 per group) charged with developing minimum proficiency standards.

Method
Applied Nedelsky procedure to reading and math tests for grades 3, 6, 9, and 11. Used discussion to describe rationale behind identifying distractors minimally competent examinees could easily eliminate as incorrect.

Results
Nedelsky resulted in highly discrepant standards across grade level and content area (e.g., 52.5%-81.9% for reading, grades 3, 6, 9, and 11; and 58.2%-65.9% for math, grades 3 and 6). For math grade 9, 37.2%; and math grade 11, 37.3%.

Conclusions
1. Judges not comfortable with math grade 9 and 11 cutoffs set with Nedelsky.
2. No substantial agreement or pattern of disagreement between cutoffs developed by Nedelsky and Contrasting Groups.
3. Recommends no one standard setting procedure be relied on to determine cutoffs.

** Includes formulas for Linear Discriminant Function (LDF) and Quadratic Discriminant Function (QDF) for minimizing errors of misclassification (i.e., false positives and false negatives). Includes discussions on their use.

Recommendation
Helpful for looking at comparisons of judgmental and empirical standard setting procedures. Serves as a source for LDF and QDF formulas for minimizing false positives and false negatives.

Mills, C. N. (1983). A comparison of three methods of establishing cut-off scores on criterion-referenced tests. Journal of Educational Measurement, 20, 283-292.

Hypothesis/Goal
Most studies which compare standard setting methods use different judges to obtain standards from different methods. Discrepancies in standards obtained in these studies may be due to method differences, but may also be due to differences in groups of judges.

Present study compares Angoff, Borderline Group, and Contrasting Groups procedures. Two groups use Angoff procedure to set standards. One group previously used Borderline Group method; other group previously used Contrasting Groups method.

Participants
Second grade teachers in Louisiana.

Method
Prior to pilot test administration, teachers answered for each student the following question: "How would you expect this student to perform on a fundamental Grade 2 language arts/mathematics test?

- o I would expect this student to pass.
- o I would NOT expect this student to pass.
- o I would be unable to predict this student's performance as clearly passing or failing" (p. 285).

After pilot test administration, a sample of these teachers set standards using the Angoff procedure on the test representing their content area (e.g., language arts teachers set standards on the language arts test, math teachers on the math test).

Results
1. Borderline Group resulted in highest standards.
2. Angoff and Contrasting Groups methods were most similar.

Conclusions
1. Borderline Group method may have resulted in such widely discrepant cutoffs compared to other methods because of instructions for identifying borderline students. Borderline students were those whose performance the teacher couldn't "predict." Clearly competent and non-competent students may have been classified as borderline merely because the teacher did not have enough information on which to base a judgment (e.g., student recently transferred into new class).
2. Angoff standards will be similar to Contrasting Groups standards when same judges are used for both methods.

Recommendation
Should be considered further if we use the same judges to set standards with different procedures and if we want to examine the dispersion of standards set this way. Should be considered when comparing and contrasting standard setting procedures.

Norcini, J. J., Lipner, R. S., Langdon, L. O., & Strecker, C. A. (1987). A comparison of three variations on a standard-setting method. Journal of Educational Measurement, 24, 56-64.

Test of gastroenterology subspecialty of internal medicine divided into two matched halves - matched in content and psychometric properties. Test administered to gastroenterology specialists (GI) and general internal medicine practitioners.

Before study began, SMEs received article describing Angoff standard setting method. Group meeting held to discuss Angoff method, describe borderline group (those who would barely pass or fail internal medicine certification exam), and practiced using method.

Later, SMEs were mailed instructions and booklet containing items from first half of test. Indicated for each item: correct answer, proportion of examinees that answered correctly, and whether examinees were GI or general internal medicine. Instructed to pay close attention to whether examinees were GI or general. SMEs estimated percentage of borderline examinees who would answer each item correctly. Judgments were made independently.

At later group meeting, SMEs discussed borderline group characteristics and their ratings on items from first half of test. SMEs with highest and lowest ratings stated their rationale for their ratings. After discussion and consulting their own independent ratings, SMEs rated items again. SMEs not allowed to alter initial ratings.

A month later, SMEs received instructions and booklet containing second half of test. Instructions similar to those mailed initially.

Cut off scores set before, during, and after meeting not statistically significantly different. Ratings collected during and after group meeting were more reliable than those collected before meeting.

Recommendation

Deserves more consideration if we use a survey approach (i.e., mail instructions vs. a workshop setting with workshop leader).

Plake, B. S., & Melican, G. J. (1989). Effects of item content on initial judge consistency of expert judgments via the Nedelsky standard setting method. Educational and Psychological Measurement, 49, 45-51.

Hypothesis/Goal

Propose alternative Nedelsky procedure - consider items individually as part of an item pool. Later, develop a test by selecting items from pool. Minimum Passing Levels (MPL's) associated with selected items are used to set the standard. Assume: (a) "minimal competency" definition remains constant from the generation of items for an item pool to the selection of items for a test, (b) MPLs not based on test form contextual variables (i.e., difficulty, content, length).

Examine robustness of judgments made using Nedelsky method when items reviewed under different test form contexts and over relatively long period of time between ratings. Test form contexts investigated - test length and overall test difficulty.

Method

5 experts used Nedelsky method to set standards on 48-item math test with mean item difficulty of .40. Test used to identify students who need math remedial help.

1 year later, 28-item math test with mean item difficulty of .65 was developed. Consisted of 15 items from earlier test. Same 5 experts used Nedelsky method to set standard.

Results

- For 48-item test, cutoff for 15 items was 7.35, standard deviation = 2.35.
- For 28-item test, cutoff for 15 items was 6.36, standard deviation = 2.17.
- To assess agreement of actual item ratings of 15 items across review occasions, number of incorrect alternatives predicted to be eliminated by minimally competent examinee for each item was correlated for each judge. Average correlation was .55, standard deviation = .22
- To evaluate consistency of judgments across occasions (i.e., were same distractors eliminated both times, developed consistency index). CI = .67 or experts were consistent across occasions for 67% of their decisions.

Conclusions

Judges were fairly consistent in assessments using Nedelsky method on same items regardless of test length or overall test difficulty. Developing banks of items evaluated using Nedelsky method may be viable approach for setting cutoffs for exams developed from item bank.

Comment

Not helpful for our project because we're not developing a test from an item bank.

Poggio, J. P. (1984). Practical considerations when setting test standards: A look at the process used in Kansas. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA, April, 1984.

Paper presents summary of what learned from using different methods to set standards on Kansas minimum competency tests from 1980 - 1984.

Used judgmental (Angoff, Ebel, Nedelsky) and empirical (Contrasting Groups, Borderline Group) methods. Two formats for judgmental methods: (a) convene panels of judges and (b) mail survey questionnaires to judges.

Two general observations: (a) "no one method has surfaced as the one to use that identifies the true cut-score" and (b) "comparisons of the cut-scores over methods is altogether consistent with other research represented in this area" (pp. 2-3).

Empirical Methods: Contrasting Groups and Borderline

- o Rather easily implementable
- o Teachers report little difficulty in following what is to be done in Contrasting Groups
- o Borderline method creates some confusion, i.e., "What do you mean just barely minimally competent?"
- o Standards tend to be lower than for Angoff or Ebel
- o Standard becomes available well after actual testing
- o Public is confused and tends to doubt legitimacy of standard when they can't understand the "statistical magic" that produces standard
- o Methods give support to contention that "teachers can already tell us who is competent"

Judgmental Method: Nedelsky

- o Judges report being very confused not confident in their judgments
- o Can be used only by experienced teachers
- o Judges tend not to be careful in studying items and often mark the correct choice as being not a viable distractor
- o Standard is substantially lower than all other methods; therefore, data from it is quickly ignored

Judgmental Method: Angoff

- o Easy to implement and understand either in panel or survey format
- o Judges tend to establish their own "mean" level causing considerable variability among individual judge standards. Becomes particular problem in panel approach when few judges are used
- o Defining students "who are minimally competent" is problem for many judges

Judgmental Method: Ebel

- o Task itself is time consuming. Fatigue and boredom can become problem
- o Easy for most judges to understand and can be implemented without difficulty
- o Relevance rating of "Questionable" causes judges to become concerned about the method

- o Cell percent passing task causes real difficulty/debate over "Questionable" dimension
- o Standard can very considerably depending on whether it's computed by judge or based on group cell values

Conclusions:

- o Nedelsky and empirical methods no longer used
- o Use both Ebel and Angoff methods
- o Use survey approach because: (a) more efficient relative to time and cost, (b) permits broader base for input into decision-making process, (c) standards across Ebel and Angoff are comparable and psychometrically favor survey approach
- o "Yet, once the data are obtained the actual setting of each test's standard is not solved by the mathematics prescribed by the method. In fact, it is interpolated for the data gathered by a 26-member State Advising Committee" (p. 5).
- o "The process, while objective to a point, remains largely value laden" (p. 5).

Recommendation

Provides excellent summary of standard setting methods used in Kansas minimum competency program, which turns out to be almost all traditional standard setting methods. Pros and cons raised are similar to those found in other studies. At least here, pros and cons are all in one place.

Poggio, J. P., Glasnapp, D. R., & Eros, D. S. (1981). An empirical investigation of the Angoff, Ebel, and Nedelsky standard setting methods. Paper presented at the Annual Meeting of the American Educational Research Association, Los Angeles, CA, April 1981.

Hypothesis/Goal
Simultaneously compare Angoff, Ebel, Nedelsky, and Contrasting Groups for 10 different tests (reading and math for grades 2, 4, 6, 8, and 11) in Kansas minimum competency testing program.
Look at discrepancies among standards, reliabilities and validities of standards.

Participants
Kansas school districts. Participation voluntary.

Method
Each district assigned Angoff, Ebel, or Nedelsky. Each district set standards for 6 of 10 tests. Assignment of tests was random but ensured approximately equal number of ratings for each test. District coordinator selected judge from appropriate content area (reading, math) and grade level for each of the 6 tests assigned to district. Judges were to have at least 2 yrs teaching experience in applicable content and grade level. Judges received packets of material including instructions from the district coordinator. Upon completion, packets were returned to the district coordinator.

Results
1. All methods resulted in a slightly negatively skewed distribution.
2. Angoff cutoffs contained more variability than Ebel or Nedelsky.
3. Ebel consistently resulted in higher cutoff. Angoff resulted in cutoff in same region of distribution but generally 1-5 points lower. Ebel and Angoff resulted in substantially higher cutoffs than Nedelsky.
4. Reliability (ANOVA coefficient alpha analyses) high for Angoff, Ebel, and Nedelsky (r = .89 was lowest coefficient).

Conclusions
1. Nedelsky yields lowest raw score standard, Contrasting Groups next lowest, Angoff next, and Ebel yields highest.
2. Order is consistent for given test.
3. Contrasting groups minimizes number of misclassification errors. Nedelsky results in more false positives. Angoff and Ebel increase false negatives.
4. Ebel and Angoff most similar in cutoff, but Angoff cutoffs have more variability.

Comment
Excellent comparison of 3 judgmental standard setting procedures. It serves as a basis for Poggio, J. P. (1984). Practical considerations when setting test standards: A look at the process used in Kansas. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA, April, 1984, and a more comprehensive summary is found in that review.

Popham, W. J. (1978). As always, provocative. Journal of Educational Measurement, 15, 297-300.

Rejoinder to Glass, G. V. (1978). Standards and criteria. Journal of Educational Measurement, 15, 237-260.

Argues that while all standards setting methods require some subjective judgments these judgments are not made in a vacuum. "There are myriad instances in which people exercise their judgmental powers in a decisively nonarbitrary fashion" (p. 298) (e.g., judging exotic wines, popularity of musicians, etc.).

That different procedures incorporating different elements intended to accomplish same function yield divergent results is not surprising (e.g., different methods of calculating taxes). Must decide which procedure is more appropriate.

Two conceptions of minimal competency: (a) "requisite for the future" - absolute minimum level of skill needed to function effectively in school, society, etc. and (b) lowest level of proficiency considered acceptable for situation at hand.

Discusses what seems to be a contradiction in using normative data to set standards on criterion-referenced test. Popham argues that using normative data to set standards does not compromise underlying assumptions of criterion test developers.

Standard setting methodology needs to be improved, however, one should not discard the notion as hopeless.

Comment
As rejoinder to Glass, it demonstrates just how volatile standard setting issue can be.

Popham, W. J., & Husek, T. R. (1969). Implications of criterion-referenced measurement. Journal of Educational Measurement, 6(1), 1-9.

This paper examines the differences between norm- and criterion-referenced measurement in terms of variability, item construction, reliability, validity, item analysis, reporting, and interpretation.

Variability - Said to be at the core of the difference between norm-referenced (NR) and criterion-referenced (CR) tests. The meaningfulness of a NR test is dependent on the relative position of the score in comparison with other scores, the more variability the better. With CR tests variability is irrelevant. In CR tests meaningfulness comes directly from the connection between items and the criterion. Variability is not a necessary condition for a good CR test.

Item Construction - During item construction a NR test writer is concerned with variability so "too easy" or "too hard" items are avoided and wrong option answers are increased. The CR item writer is more concerned with the accurate reflection of criterion behavior. A second difference is while in both NR and CR tests equivalent forms are required when assessing individuals; when assessing programs equivalent forms are not needed.

Reliability - If a CR test is tied to the criterion it should be internally consistent and the items should be similar in terms of what they are measuring. Assessing internal consistency is difficult, however, since it is related to variability classic indices of reliability may not be appropriate.

Validity - Procedures for assessing validity in NR tests are based on correlations and thus on variability. CR measures are validated primarily in terms of the adequacy in which they represent the criterion. Content validity approaches are more suited to such tests.

Item Analysis - NR measures use item analysis to identify those items which are not properly discriminating (i.e., too hard, too easy, ambiguous) among individuals taking the test. With the use of CR tests, discrimination indices must be modified. If an item reflects an important aspect of the criterion yet does not discriminate among individuals then it need not be eliminated.

Reporting and Interpretation - With respect to NR methods since performance is measured with respect to performance of other individuals group-relative descriptors are used such as percentile rankings or standard scores. When interpreting individual performance with the use of CR measures such group-relative descriptors are not appropriate.

Different Kinds of Criterion-Referenced Tests - There tend to be two types of CR tests; one is ideal, the other more typical. In the ideal case items are tied to the criterion and the test is homogeneous in a special sense. The meaning of a test score is thus altogether unambiguous. Those individuals obtaining the same test score received it in the exact same manner. The second type of CR test is more typical of today's testing. The items on a test form a sample of a potentially larger group of items generated from a criterion. The test score is not completely unambiguous in that it is not

known from the test score which items an individual missed. Yet the score does allow for approximations of criterion behavior.

Pulakos, E., Wise, L., Arabian, J., Heon, S., Delaplane, S. K. (1989). A review of procedures for setting job performance standards. Washington, DC: U.S. Army Research Institute for the Behavioral and Social Sciences.

Much research on setting minimum performance standards in educational testing and professional licensure, but little or none on job performance measures. Much research on setting standards on written, especially multiple-choice, tests, little or none on worksample or hands-on tests.

Job performance standards useful for following personnel activities:

1. Employee motivation
2. Identifying training needs
3. Evaluating personnel programs
4. Setting minimum entry standards

Army's primary interest is to use job performance standards as basis for setting selection standards, i.e., link predictor scores to job performance levels.

As part of Job Performance Measurement project (Project Alpha), developed several job performance measures:

1. Hands-on tests for sample of 15 tasks scored in terms of percentage of steps performed correctly
2. Multiple choice job knowledge tests covering 30 tasks including 15 tasks tested hands-on
3. Supervisor and peer ratings of 11 army-wide performance dimensions and 7-12 MOS specific dimensions
4. Administrative records of performance (e.g., awards and certificates, disciplinary problems, PT scores)
5. Supervisory simulation exercises (personal discipline, counseling, and training) for second tour only

In all MOS, need distribution of quality soldiers. Some percentage should be at least minimally competent, and some percentage of higher ability. Minimally competent OK for entry positions, but higher ability needed for strong NCO corps.

Army needs procedures that:

1. Can be applied to Project A criterion measures
2. Yield reliable, multiple performance standards
3. Indicate how standards reflecting multiple dimensions of performance should be combined into overall standard
4. Provide mechanism for linking performance standards to selection standards

Judgmental Paradigms - describes standard setting methods. Discusses advantages and disadvantages of each and how each may be applicable to Army's standard setting situation. (Because methods and their empirical evaluations are outlined elsewhere, following summary outlines applicability to Army's situation only.)

1. Item-Based Methods (in other literature may be referred to as Judgmental Methods)
   A. Nedelsky - can be used for multiple choice knowledge tests only. Impractical for setting multiple performance standards. Problems defining "minimally competent examinee" (minimally competent on specific task or in MOS in general)
   B. Angoff - can be used for dichotomously scored items. Adaptation required for continuous measures.
   C. Ebel - restricted to dichotomously scored items. Too time consuming. Likely to lead to stricter standards than desired. Criterion measures may not yield themselves to difficulty/relevance stratification
   D. Jaeger - can be applied easily to written or hands-on tests. May yield stricter standards than desired. Adaptation required for continuous items.

2. Examinee-Based Methods (in other literature may be referred to as Empirical Methods)
   A. Borderline-group
   B. Contrasting-group
   C. Both can be used with dichotomous or continuous items. Army supervisors accustomed to making examinee-based judgments. Can't be used for new MOS. Appropriate referent population must be identified. May be hard to find "unacceptable" examinees.

3. Outcome-Based Methods (in other literature may be referred to as Decision Theory or Utility Theory) - originally designed to assist in making decisions from among several alternatives (e.g., selecting new employee from among several applicants, selecting among selection devices). Using Decision Theory to make binary decisions (e.g., whether to retain, promote, retrain) requires some extension of general Decision Theory paradigm.

4. Other Methods -
   A. Berk (1976) - Use empirical data of "instructed" and "uninstructed" examinees. Set standard three ways: (a) classify outcome probabilities, (b) compute validity coefficient, (c) utility analysis. Fairly easily understood. Problem equating "instructed" with "competent."
   B. Kriewall (1972) - Classify examinees as non-master, master, or in-between. Model focuses on identifying likelihood of classification errors and requires several assumptions (e.g., randomly selected dichotomously scored items, independent responses to items). Set boundary values, decide on initial cutoff, and estimate misclassification errors based on cutoff. Actual data not needed. Requires satisfying several assumptions and is complicated. May not be suitable for Army use.
   C. Cangelosi (1984) - Set cutoff as develop test objectives. As develop objectives, specify proportion of correct answers. Borderline examinee expected to achieve for items representing each objective. Cutoff is weighted sum of expected proportions for all

objectives. If define "success" early in test development process, test may be more valid. Very difficult to implement. Yields highly inconsistent and strict standards.

Judgment Process

1. Judgment Facilitation Techniques
   A. Normative Data - likely to serve as reality check mechanism and help increase consensus among judges.
   B. Iterative Judgment Process - if data on standards simply presented, probably lead to shift in ratings toward group mean, median, or mode. If judges allowed to state rationale for their standards, must ensure that dominant judge doesn't inappropriately influence group.

2. Judge Characteristics - should include all "interest" groups, i.e., NCOs, Officers, TRADOC, FORSCOM. Should be experts. For item-based methods, judges should be knowledgeable about distribution of examinees on measure of interest. For examinee-based methods, judges should be knowledgeable about actual job performance of soldiers being classified. Should not include test developers (standards will be too strict).

3. Judge Training - familiarize judges with test for which will be setting standards. Familiarize judges with standard setting procedure to be used. How to interpret normative data.

4. Number of Judges - no clear guidance in literature. Too few judges may lead to large standard error of recommended standard. Too many may waste resources and prolong or complicate standard setting process.

Combining Multiple Standards - given that job performance is multidimensional, how can you combine them to yield overall standard? Range from simple linear composite to conjoint measurement (tradeoffs in good and bad performance among dimensions).

1. Multiple Hurdles Model - no amount of increment in other areas can compensate for below standard performance on any other dimension. If fail standard on one dimension, fail standard on overall performance.
2. Compensatory Model - decrement on one dimension could be compensated for by increments in other areas.

Army currently uses combination of Multiple Hurdles and Compensatory - if pass moral and physical screen (Multiple Hurdles), can take ASVAB composed of composites which are scored according to Compensatory Model.

Linking Selection Standards to Performance Standards - two issues: (a) lack of perfect prediction using available predictors (b) interaction with training effects (see pp. 40-43 for more detail).

Basic information for linking predictor and criterion standards:

1. Performance standards
2. Estimates of population distributions for predictor and criterion
3. Empirical or synthetic estimates of validity of selection composite to

be employed

Comment
Extremely helpful, especially because many of the Project A criterion measures are similar to AFHRL/JPM criterion measures. Also helpful because addresses issues that are peculiar to military institutions.

Sadacca, R., White, L.A., Campbell, J.P., DiFazio, A.S., & Schultz, S.R. (1989). Assessing the utility of MOS performance leveles in Army enlisted occupations. Washington, DC: U.S. Army Research Institute for the Behavioral and Social Sciences.

Abstract

Project A is the Army's long-term program to develop a complete personnel system for selecting and classifying all entry-level Army enlisted personnel. The utility measurement component deals with determining the relative utility to the Army of different levels of performance in entry-level military occupational specialties (MOS). Because little research has been performed on such questions, exploratory work was done in a series of workshops with army officers on how performance levels should be defined, what unit of measurement is appropriate for describing the relaive value of differential job assignments across various MOS/performance level combinations, and how such values can best be estimated. Two scaling methods (pile placement and direct estimation) were selected and used to estimate utility values for 273 entry-level MOS. The research established that a coherent, reliable set of relative utility values can be derived at all performance levels for a wide variety of MOS.

Sands, W. A. (1973). A method for evaluating alternative recruiting-selection strategies: The CAPER model. Journal of Applied Psychology, 57, 222-227.

Hypothesis/Goal
Demonstrate Cost of Attaining Personnel Requirements (CAPER) model, which is designed to evaluate alternative recruiting and selection strategies. Specifically, determines optimal strategy for minimizing estimated total cost of recruiting, selecting, inducting, and training sufficient number of people to meet specified quota of satisfactory personnel. Also considers cost of false positives and false negatives.

Method
Hypothetical recruiting-selection problem used to illustrate application of model. Must specify: quota, base rate, proportion of previous graduates and failures who would have qualified for acceptance at each possible cutoff (can be estimated from usual statistics, i.e., mean, standard deviation, correlation, if assume bivariate normal distribution). Also must specify following costs: recruiting, selection, induction (processing), training, erroneous acceptance, erroneous rejection.

Compared currently used selection procedure (medical clearance) and proposed new procedure (medical clearance and aptitude test). Aptitude test scores available for large sample of applicants previously admitted to program, but scores not used for selection nor were scores available to instructor.

Gives and works through equations for estimating:
1. Number of applicants who must be recruited in order to meet quota.
2. Number of erroneous acceptances.
3. Number of erroneous rejections.
4. Number of people who will be accepted.
5. Total cost of employing ordinary selection procedure to meet quota.
6. Total cost of employing experimental selection procedure to meet quota.

Total cost equations can be broken down to show costs of: recruiting, selection, induction, training, and erroneous decisions.

Results
Provides information derived from above equations for each cutoff.

Conclusion
"The CAPER model is designed to provide useful planning information to managers of personnel systems, not to replace them, nor relieve them of the responsibility for sound decision making" (p. 226).

Most important advantage - ease of communicating results. Results are presented in terms of number of people and dollar costs, which are easily understood.

User's manual including FORTRAN program and detailed documentation prepared for CAPER model (see Sands, W. A. (1971). Application of the cost of attairing personnel requirements (CAPER) model. (Tech. Bull. WTB 72-1) Washington, DC: Naval Personnel Research and Development Laboratory).

Use of model assumes that all graduates are equally useful in terms of actual on-the-job performance. Also assumes that predictor-criterion relationship is stable, i.e., base rate and experimental variable frequency distributions for graduates and failures are based on relatively large, representative sample of selectees.

Recommendations

This could be very helpful, especially if we want to estimate the cost of various cutoffs. The mathematics are straightforward. The only hard part may be coming up with estimates of the costs required by the model.

Shepard, L. (1980). Standard setting issues and methods. Applied Psychological Measurement, 4, 447-467.

Always error associated with selection of cutoff scores. Performance of individuals immediately on either side of the cutoff score will be indistinguishable from each other. With psychometrically sound test, valid distinctions can be made between individuals who score well above or well below cutoff, but "pass fail distinctions near the cutoff will have poor validity because a continuum of performance has been 'arbitrarily' dichotomized."

Continuum Standard Setting Methods - ability being assessed is assumed to be continuous and linearly related. Cutoff score is needed because dichotomous decision is needed.

State Standard Setting Methods - mastery is assumed to be all-or-none. Examinee either has skill or doesn't. Implies cutoff score of 100%, which is unrealistic given measurement error and human error (e.g. fatigue, carelessness, etc.).

To reduce variability in standards: (a) use people with different value positions and areas of expertise, (b) have judges discuss differences in ratings.

Absolute standards - allow everyone to pass if everyone is competent and no one to pass if no one is competent. This is rarely the case, e.g., consider cost of remedial education if lots of people fail. Lack of qualified applicants if lots of people fail. In such cases, standards are raised or lowered to accommodate organization's needs.

Comment
Excellent summary of issues and problems underlying standard setting.

Shikiar, R. & Saari, L. M. (1985). Establishing cut scores for the NRC reactor operator and senior reactor operator exam. (Technical Evaluation Report No. PNL-5131). Seattle, WA: Pacific Northwest Laboratory.

Norm-Referenced Tests - make inferences about test score by comparing it with distribution of test scores obtained by specific sample of test takers (e.g. SAT, GRE, grading "on a curve"). Criterion-Referenced Tests - make inferences about test score by comparing it with established standard which is grounded in content-domain measured by test (e.g. state driver's test - score above certain point means you get license, score below means you don't).

Judgmental Approaches to Setting Cut Scores on Criterion-Referenced Tests

- Angoff (1971) - Judges rate probability (0 to 1) that minimally competent individual would answer each question correctly. Sum of ratings across all items = cut score for that judge. Average of each judge's cut score = cut score for test.

- Nedelsky (1954) - Given multiple choice format, judges examine each alternative for each item. Eliminate all alternatives that minimally competent individual would recognize as incorrect. Cut score for given judge = sum of items as weighted by judgments on item alternatives. Most widely used method for setting cut scores for licensure exams.

- Ebel (1972) - Judges sort items by item difficulty and item relevance then review all items in given cell and judge proportion that should be answered correctly by minimally competent individual. Product of this proportion and number of items in each cell is summed for all cells. Sums are averaged across judges = cut score.

External Test Information Approaches to Setting Cut Scores on Criterion-Referenced Tests

- Contrasting groups method - Administer test to two groups - one judged to be competent and one judged to be non-competent. Compare distribution of two groups. Cut off score selected to maximally differentiate between two distributions.

- Borderline group method - Administer test to group judged to be borderline between competent and non-competent. Mean of distribution judged to be cut score.

Rather than change cut score, change test. Set cut score at 80, and develop test such that competent individuals score 80 or above and incompetent individuals score below 80. Would be politically feasible - cut score of 95 or 35 would not set well with most people. Teachers currently write tests according to the 90, 80, 70, 60 grading scale.

Comment
It's not the most helpful study available. It's interesting because standards were set in a non-educational setting.

Skakun, E. N., & Kling, S. (1980). Comparability of methods for setting standards. Journal of Educational Measurement, 17, 229-235.

Hypothesis/Goal

1. Introduces modification of Ebel.
2. Investigates whether Nedelsky, modified Ebel, or currently used normative approach generate similar passing scores on national certifying exam in General Surgery.
3. Reports effects of different passing scores on overall pass rate.
4. Investigates reliability of mean ratings given to items and variability of judges' passing scores for each absolute standard approach.

Participants

8 judges who had been actively involved in General Surgery Test Committee with writing multiple choice test items and other aspects of testing. 7 judges used Nedelsky, 6 used each of two Ebel methods, and 5 were common to all 3 procedures.

Method

194 items from General Surgery test item library. Each item classified by writer on relevance (essential, important, acceptable) and taxonomy (factual, comprehension, problem solving). From prior administrations, item difficulty data already available.

Ebel I method required classifying items on difficulty and taxonomy. Ebel II required classifying items on relevance and taxonomy. Because items already classified, judges merely review items in cells and indicate proportion of items that "barely qualified" candidate would be expected to answer.

Currently used normative approach - cutoff is score that's 1 standard deviation below mean for Reference Group (first-time test takers who were educated at approved North American training program). Apparently it's possible for standard to change across groups of examinees.
Nedelsky completed first, 6 months later Ebel I, 3 days later Ebel II.

Results

Nedelsky produced most dispersion in cutoffs. Ebel I and II about the same amount of dispersion.

Nedelsky was least reliable method ($\underline{r}$ = .61 compared to $\underline{r}$ = .98 for Ebel I and II).

Nedelsky produced lowest standard (66.7%), Ebel I next (69.7%), normative method next (70.6%), and Ebel II highest standards (71.7%)

Ebel II failed largest percentage of examinees (45.6%), normative method second largest (41.3%), Ebel I next (35%), and Nedelsky the fewest (22.5%).

Conclusion

Different standard setting methods produce different cutoff scores. Variability may be due to differences in judges' definitions of "barely qualified".

Wigdor, A. K., & Green, B. F. (eds.), (1986). Assessing the performance of enlisted personnel: Evaluation of a joint-service research project. Washington, DC: National Academy Press.

Major Themes of Report: It is the committee's considered opinion that the assessment of overall job competency or job mastery ought to be of central concern in the development of performance measures. The committee also feels that individual normative comparisons should be replaced with some indicator of absolute performance. A second theme concerns the need for close attention to the problems of standardizing the administration of hands-on job-sample tests. A third major theme is the need to strengthen the theoretical structure of the research. To improve understanding of what is being measured, the research must involve a vigorous, continuing process of hypothesis testing.

Overview of the Report and Recommendations

Chapter 1 reviews the basic design of the Joint-Service Project, briefly describes what each of the Services has done to date, and presents the committee's overall evaluation of the research.

Recommendation: That the Joint-Service Project should go forward.

Chapter 2 proposes that the design of the Joint-Service Project be expanded to embrace the measurement of overall job competence, so that test scores reveal how well individuals perform with reference to the whole job.

Recommendation: The conceptual foundations of the Joint-Service Project should be enlarged to embrace the measurement of overall competence levels in addition to measuring performance variations among individuals.

The committee looks at the three stages in the development of criterion measures--job analysis, the selection of tasks, and the development of scoring strategies. Recommendations are provided for each of the stages and can be found in the text.

Chapter 3 briefly summarizes the scientific argument for standardizing the administration of tests and performance measures and then presents the observations concerning standardization made by committee members during the administration of Army, Air Force, and Navy hands-on and written tests. The committee's general conclusion is that there are a number of serious threats to standardization in the administration of hands-on tests that could seriously compromise the Joint-Service Project and recommendations are made for the reduction of threats to standardization of administration in terms of Logistics, and Selection and Training of Test Administrators.

Chapter 4 discusses the analytical challenges that the Services face in evaluating the comparative adequacy of criterion measures under development: hands-on tests, interviews, simulations, written job knowledge tests, and rating scales.

Recommendation: For at least one job, it would be highly desirable for each Service to develop and administer all five types of criterion measure,

possibly borrowing and adapting from the work of the other Services, to obtain direct evidence of the comparability of the measures.

Wood, R., & Power, C. (1987). Aspects of the competence-performance distinction: Educational, psychological and measurement issues. Journal of Curriculum Studies, 19(5), 409-424.

This paper is an investigation of the competence-performance distinction. The authors point out that it had become apparent in the literature that competence was being used in two recognizably different ways. There was competence as enhanced performance and competence as the deep structure responsible for the surface performance. The distinction was then regarded in a different fashion: "when we study competence we study what turns out to be an embryonic working model of the development of expertise; when we study performance we study the methodological problems common to all educational and psychological measurement in which authentic expression of what a person is really able to do or really believes or really thinks is frustrated."

The competence-performance distinction

The authors view competence in line with Messick's (1984) interpretation: "Competence refers to what a person knows and can do under ideal circumstances, whereas performance refers to what is actuallly done under existing circumstances. . . . Thus, a student's competence might not be validly revealed in either classroom performance or test performance because of personal or circumstantial factors that affect behavior." The authors view is that there is every reason not to take performance at face value.

Objections to competence theory - One objection concerns how to get from performance to statements about the unobservable competence. The difficulty in inferring competence from performance is made worse by an empirical fact--correlations across tasks ostensibly measuring the same competence or across the same task over time are often low. Inconsistency has been a problem for all structural competence theories.

Elaborative procedures versus training - In order for the individual being tested to optimize performance, there are those who would propose to develop training tasks to encourage the manifestation of latent competence in overt performance.

Inferring competence from performance - In assessing performance on tests of underlying competence, it must be established that there is only one source of failure (a lack of competence) and only one route to success (that deriving from the appropriate competence).

Hints - A rather special type of elaborative procedure in that a hint is only useful to those who possess a sufficient amount of a competence to make sense of the hint.

The nature of competence - The authors suggest competence is the product of some education, training or other experience, rather than being an inborn or natural characteristic. Job competence involves the application of knowledge and skills, and is exhibited through purposeful and real, rather than simulated activity.

Development of competence - The development of competence seems to be contingent on the simultaneous development of general cognitive abilities through learning and transfer from a variety of developmental contexts (school, home, etc.), and of special knowledge structures, generally through systematic instruction.